THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

A correction for regression discontinuity designs with group-specific mismeasurement of the running variable

Citation for published version:
Bartalotti, O, Brummet, Q & Dieterle, S 2020, 'A correction for regression discontinuity designs with group-specific mismeasurement of the running variable', *Journal of Business & Economic Statistics*, vol. N/A, pp. 1-16. https://doi.org/10.1080/07350015.2020.1737081

Digital Object Identifier (DOI):
10.1080/07350015.2020.1737081

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Business & Economic Statistics

Publisher Rights Statement:
This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of Business & Economic Statistics on 3/3/2020, available online: https://www.tandfonline.com/doi/full/10.1080/07350015.2020.1737081.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 08. May. 2023

A Correction for Regression Discontinuity Designs with Group-Specific Mismeasurement of the Running Variable

Otávio Bartalotti, Quentin Brummet, and Steven Dieterle*

February 4, 2020

Abstract

When the running variable in a regression discontinuity (RD) design is measured with error, identification of the local average treatment effect of interest will typically fail. While the form of this measurement error varies across applications, in many cases the measurement error structure is heterogeneous across different groups of observations. We develop a novel measurement error correction procedure capable of addressing heterogeneous mismeasurement structures by leveraging auxiliary information. We also provide adjusted asymptotic variance and standard errors that take into consideration the variability introduced by the estimation of nuisance parameters, and honest confidence intervals that account for potential misspecification. Simulations provide evidence that the proposed procedure corrects the bias introduced by heterogeneous measurement error and achieves empirical coverage closer to nominal test size than "naive" alternatives. Two empirical illustrations demonstrate that correcting for measurement error can either reinforce the results of a study or provide a new empirical perspective on the data.

Key Words: Nonclassical Measurement Error, Regression Discontinuity, Heterogeneous Measurement Error.

JEL Codes: C21, C14, I12, J65

*Bartalotti: Department of Economics, Iowa State University and IZA. 260 Heady Hall, Ames, IA 50011. Email: bartalot@iastate.edu. Brummet: NORC at the University of Chicago, 55 East Monroe Street, 31st Floor, Chicago IL 60603. Email: brummet-quentin@norc.org. Dieterle: School of Economics, University of Edinburgh, 31 Buccleuch Place, Edinburgh, United Kingdom EH8 9JT. Email: steven.dieterle@ed.ac.uk.

1 Introduction

Regression Discontinuity (RD) designs have become a mainstay of policy evaluation in many social science fields. These designs rely on treatment assignment being based on a "running variable" passing a particular cutoff, which is observed by the researcher. In practice, however, there are multiple forms of measurement error in the running variable that, when present, will invalidate this approach.

We consider situations in which a researcher has access to data with the running variable exhibiting group-specific measurement error, where each group faces potentially different measurement error distributions. This encompasses a wide range of situations, as the group's definition can be tailored to the specific knowledge/beliefs of the researcher regarding the heterogeneity in the measurement error. In some cases, the definition of the group structure will follow naturally from the data available and the setting being studied. In other cases, this definition may be subject to more discretion on the part of the researcher, but our setup makes it possible to consider the robustness of the results to plausible, alternative group definitions. Our approach is best suited for a situation in which the researcher can learn about the mismeasurement using auxiliary data, but could also be implemented without auxiliary data if the researcher has knowledge about the measurement error distribution for each group.

One prominent example identified by Barreca et al. (2011) (hereafter, BGLW) is the "heaping" of birth weight measures at particular values due to hospitals using scales with different resolutions. In this setting, additional care is given to babies born at a birth weight of strictly below 1500 grams, allowing for an RD analysis of the effect of the additional resources on child outcomes. However, some hospitals record the weight to the nearest gram while others record it at ounce or gram multiples— 5g, 10g, and up to 100g multiples. Therefore, the treated units measured at 1499g are likely to be well measured and accurately reflect the mean outcomes and unobservables at the true weight of 1499g, but the closest untreated units measured at 1500g will reflect the mean outcomes and unobservable factors for babies with a true weight up to 50g away. The problem is further complicated by the nearby ounce multiple measure at 1503g that will have a much different measurement error distribution than babies measured at 1500g. Depending on the gradient of child outcomes with respect to the true birth weight, this could generate a spurious discontinuity at the cutoff of the mismeasured running variable.

Another example comes from geographic regression discontinuity (GeoRD) settings, where the running variable is often measured as the distance from an individual's residence to a border

that separates two policy regimes.[1] Ideally, researchers would use a precise distance measure from the residential address to the border. However, due to data limitations it is common to use the distance from the geographic centroid of a larger region to calculate the distance to the border. Due to differences in region size and population distribution within each region, units in different regions— or groups— will face different measurement error distributions. Importantly, the centroid measure may be closer or farther away from the border than the true distance for many of the units— again creating the possibility of a discontinuous jump in unobservable factors at the cutoff in the measured distance. In this case, defining the measurement error groups by the reported regional location for units is quite natural.

Our procedure leverages auxiliary information about the measurement error distributions for various groups (see Hausman et al. (1991); Lee and Sepanski (1995); Chen, Hong, and Tamer (2005) and Davezies and Le Barbanchon (2017) for other approaches using auxiliary information to address measurement issues). This information is used to transform the observed data, re-centering the observed running variable around the moments of the underlying latent running variable distribution for each observation or group. Intuitively, this re-centering procedure corrects the distortions caused by the measurement error since, on average, some observations will be closer or farther from the cutoff than the observed running variable would indicate. The re-centering identifies the parameters on the conditional expectation of the outcome with respect to the true (unobserved) running variable rather than the mismeasured one.

The measurement error correction procedure's implementation requires the (potentially non-parametric) estimation of the moments of the multiple measurement error distributions. Hence, we develop procedures for valid inference, developing a novel asymptotic distribution approximation that considers the variability introduced by the multi-sample first stage estimation on the estimates for the ATE at the cutoff. We also extend the recent developments in Armstrong and Kolesár (2018) to provide honest confidence intervals (CIs) in the presence of measurement error in the running variable. The honest CIs allow for inference that is robust to certain mispecifications of the conditional mean function.

Importantly, the main contribution of our procedure is to address a new class of problems in which different types of measurement error affect groups of the population — such as our motivating examples above— not previously covered by the literature. Furthermore, we study both the case in which treatment is determined by the unobserved running variable, which has

[1] See Keele and Titiunik (2014) for a general discussion of GeoRD and for examples see Black (1999); Lavy (2006); Bayer, Ferreira, and McMillan (2007); Lalive (2008); Dell (2010); Eugster et al. (2011); Gibbons, Machin, and Silva (2013); Falk, Gold, and Heblich (2014).

been the focus of the majority of the existing literature, and the case in which treatment is determined based on the mismeasured running variable, which is common in relevant applications such as the very low birth weight example from Almond et al. (2010) (hereafter, ADKW) and BGLW discussed in Section 4.

These results complement the growing literature on measurement error in RD designs by considering both sharp and fuzzy designs while allowing for non-classical and heterogeneous group-specific measurement error (Lee and Card, 2008; Pei and Shen, 2017; Yu, 2012; Dong, 2015; Davezies and Le Barbanchon, 2017; Dong, 2017; Barreca, Lindo, and Waddell, 2016). For example, Pei and Shen (2017) provide sufficient conditions for identification without the need for auxiliary data in the case of discrete running variables (and under additional assumptions for continuous running variables), but only in the case with classical measurement error that is discrete and bounded, rather than more general forms.

Similar to our procedure, Dong (2015) and Davezies and Le Barbanchon (2017) both allow for non-classical measurement error. Dong (2015) focuses on the rounding case in which the measurement error distribution is homogeneous across individuals and known to the researcher, using this knowledge of the measurement error distribution to identify the ATE at the cutoff. Under homogeneous and known measurement error, such as the age rounding case studied by Dong (2015), our approach is conceptually very similar to the proposed correction from that paper. Davezies and Le Barbanchon (2017) exploit auxiliary data on the treated individuals to recover identification of the effect of interest and propose a nonparametric estimator for the continuous running variable case. Their approach has the benefit of not relying on local parametric assumptions for identification, but can only be used in "two-sided fuzzy designs." Our procedure can also be applied to cases where the measurement error can be characterized as discrete measurement of a continuous true measure, similar to Lee and Card (2008) who consider the more restrictive case where measurement error can be cast as random specification error. Importantly, our approach is the first to directly address the discontinuous change in measurement error types near the cutoff — a prominent feature of our empirical examples.

Simulation results provide evidence that our procedure performs well, successfully mitigating the measurement error induced bias and obtaining adequate test coverage in contrast to naive approaches. Most relevant to practitioners, the proposed correction improves markedly over naive alternatives even when the polynomial order is unknown and chosen using a data-driven algorithm. Moreover, our novel honest CIs provide inference that is robust to misspecification of the conditional mean function in the simulations.

In the context of the low birth weight example in ADKW and BGLW, we find that correcting for measurement error yields estimates consistent with the original results in ADKW, suggesting a large effect of very low birth weight classification. Further, estimates using our correction are much less sensitive to the exclusion of observations at "heaped" measures near the cutoff — the "Donut RD" proposed by BGLW— than the uncorrected estimates. We also apply our procedure to examine the effect of Unemployment Insurance (UI) benefit extensions during the Great Recession on unemployment studied by Dieterle, Bartalotti, and Brummet (Forthcoming). In this paper we focus on the single case of the Minnesota-North Dakota border during 2010 to highlight the intuition for our approach in the geographic setting. Here, we find that the uncorrected estimates are 18 times larger than the corresponding estimates using the moment-based correction, implying a sizable bias due to the measurement error.

The paper proceeds as follows: Section 2 presents the setup and derives the measurement error-corrected RD approach; Section 3 presents Monte Carlo evidence about the performance of the proposed method; Section 4 applies our procedure to the very low birth weight example; Section 5 applies our procedure in the GeoRD context; and Section 6 concludes.

2 Running Variable Measurement Error Correction

2.1 Setup and Motivation

Consider a basic RD setup. The interest lies in estimating the average treatment effect of a program or policy in which treatment status ($D = \{0, 1\}$) is determined by a score, usually referred to as "running variable" (X), crossing an arbitrary cutoff (c). Let $Y_1 \equiv y_1(X)$ represent the potential outcome of interest if an observation receives treatment and $Y_0 \equiv y_0(X)$ the potential outcome if it does not. The researcher's intent is usually to estimate $E[Y_1 - Y_0|X = c]$, the average treatment effect at the threshold. If observable and unobservable factors influencing the outcome evolve continuously at the cutoff then the average treatment effect at the cutoff is identified nonparametrically by comparing the conditional expectation of $Y = DY_1 + (1 - D)Y_0$ on either side of the cutoff:

$$\tau = \lim_{a \downarrow 0} E\left[Y|X = c + a\right] - \lim_{a \uparrow 0} E\left[Y|X = c + a\right]. \tag{2.1}$$

Now, consider the case in which instead of the running variable, X, we observe a mismeasured version, $\tilde{X} = X - e$, where e is the measurement error. This measurement error can be quite general including non-classical forms and can be dependent on either X or $\tilde{X}$.

The effect this measurement error has in the RD setup depends in part on whether treatment is assigned based on the true unobserved running variable or the observed mismeasured variable. As shown by Pei and Shen (2017) and Davezies and Le Barbanchon (2017), measurement error could lead to loss of identification at the cutoff in the case that treatment is defined based on the true unobserved running variable — so that $D = \mathbf{1}(x < c)$. Interestingly, if the treatment is determined based on the mismeasured running variable — implying $D = \mathbf{1}(\tilde{x} < c)$, then the traditional RD design estimates a weighted average of the treatment effect for the subpopulation for which $X + e = \tilde{X} = c$, with weights directly proportional to the *ex ante* likelihood that an individual's value of $\tilde{X}$ will be close to the threshold. This is similar to the situation described in Lee (2008) and Lee and Lemieux (2010) where individuals can manipulate the running variable with imperfect control.

Identification of the ATE is further complicated by the averaging across groups with heterogeneous measurement errors on both sides of the cutoff. Discontinuous changes in the share of groups at the cutoff introduce bias to treatment effect estimates. Measurement error correction approaches that ignore the group heterogeneity might fail to identify the intended ATE. We provide further discussion of these identification issues in Appendix A.

2.2 Assumptions

Assume that the researcher observes the treatment status D, and let $E[Y|x, D = 0] = f_0(x) + R_0, E[Y|x, D = 1] = f_1(x) + R_1$, $E[Y|\tilde{x}, D = 0] = h_0(\tilde{x})$ and $E[Y|\tilde{x}, D = 1] = h_1(\tilde{x})$; where $f_t(x)$ are polynomial approximations to $E[Y|x, D = t]$ of (potentially) unknown order J with approximation error terms given by R_t for $t = 0, 1$. For simplicity, we assume that such a polynomial is capable of capturing all relevant features in the pertinent neighborhood of the unobservable X, denoted by S.[2] We revisit this assumption in Section 2.6. Finally, let G denote the groups defining the measurement error heterogeneity, so that $\tilde{x}_{ig} = x_i - e_{ig}$, where e_{ig} is the measurement error of "type" g.

We then impose the following assumptions, which follow closely Dong (2015):

A1 $f_1(x)$ and $f_0(x)$ are continuous at $x = c$.

A2 $f_1(x)$ and $f_0(x)$ are polynomials of possibly unknown degree J, and R_1 and R_0 are negligible asymptotically in S.

[2]More precisely, let $\tilde{X} = g(X) = X - e$, then note that for a given measurement error distribution we can map any value of $\tilde{X}$ into the support of X. Let that set be $G^{-1}(\tilde{X}) \equiv \{X : X + e = \tilde{X} \text{ with probability greater than zero}\}$. Specifically, let an arbitrary neighborhood around $\tilde{X} = c$ be given by $B = [c - h, c + h]$. Then, let $A = \bigcup_{\tilde{X} \in B} G^{-1}(\tilde{x})$ be the relevant support of X and define $S = [\inf A, \sup A]$.

A3 Polynomial approximations of order J for $h_1(\tilde{x})$ and $h_0(\tilde{x})$ are identified above and below the cutoff.

A4 The main and auxiliary samples are independent from each other, and consist of i.i.d. observations of $(Y, \tilde{X}, D, G)$ and $(X, \tilde{X}, G)$, respectively.

A5 (a) For all integers $k \leq J$, $E(e^k|\tilde{x}, G = g) = \mu_g^{(k)}(\tilde{x})$, these moments exist and are identified in the support of $\tilde{X}$. (b) The conditional distribution of the measurement error for each group in the primary (p) and auxiliary (a) samples is the same, i.e., $f_p(e|\tilde{x}, G) = f_a(e|\tilde{x}, G)$. (c) The known group affiliation, denoted by G is redundant if there is no measurement error, that is, $E[Y|x, D = t, G] = E[Y|x, D = t]$.

A6 $\tilde{x}$ is redundant conditional on the true x and treatment status, i.e., $E[Y|x, \tilde{x}, D = t] = E[Y|x, D = t] = f_t(x)$ for $t = 0, 1$.

The first assumption is the usual RD identifying assumption that the potential outcomes are continuous at the threshold, so that the observed "jump" at the threshold can be associated with the causal effect of the treatment. *A*2 is a parametric functional form approximation, since local methods to eliminate the approximation error will no longer be appropriate due to the measurement error in the running variable. This assumption is flexible in the sense that it allows for a variety of approaches to approximate the conditional expectation of the outcome. For simplicity one could simply assume that $f_t(x)$ are correctly specified, implying $R_t = 0$ (Dong, 2015), or that the approximation error is mean independent so that $E[R_t|x, D = t] = 0$ (Lee and Card, 2008).

If instead one is concerned that the use of a polynomial of order J to approximate $f_t(x)$ for $t = \{0, 1\}$ will lead to misspecification bias, we show in Section 2.6 how to obtain uniformly valid inference that is robust to misspecification of the conditional mean function within a class of functions by adapting the"honest CIs" approach in Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b) to the case with measurement error. Importantly, the honest CIs provide us with an inference procedure that is consistent with recent advancements in non-parametric RD estimation allowing for misspecification, even though identification is based on a parametric assumption.

*A*3 states that we can identify a polynomial of order J that describes the mean outcome as a function of the mismeasured $\tilde{X}$. This requires $\tilde{X}$ to have sufficient variation to identify $h_t(\tilde{X})$. Our approach will exploit the mapping between $h_t(\cdot)$ and $f_t(\cdot)$ implied by separability and additivity of the measurement error to recover the treatment effect parameters. This aspect of the procedure is closely related to the approaches proposed by Hausman et al. (1991) and Dong

(2015), while Assumption $A3$ serves a similar purpose to the completeness condition required by Davezies and Le Barbanchon (2017). This assumption is more likely to hold in practice when the mismeasured running variable is continuous, and in the discrete case requires that the researcher has access to several points in the support of $\tilde{X}$ that have positive density.

In addition to the usual requirement that the data is i.i.d. and the main and auxiliary samples are independent, Assumption $A4$ implies that treatment status is observed by the researcher. This leads to different data requirements depending on the context being studied. In a sharp design, if $\tilde{X}$ is always on the same side of the cutoff as X, as in the geographic RD case analyzed in Section 5, or if treatment is determined by $\tilde{X}$, this is not restrictive at all. However, if the measurement error causes the observed running variable to cross the threshold, Assumption $A4$ requires the researcher to observe the treatment status coupled with the mismeasured running variable— perhaps in survey data where participants are asked about participation in a means tested program determined by true income falling below a certain threshold, but income (X) is only reported in discrete bins in the survey. Finally, in the fuzzy RD context, Assumption $A4$ needs to be strengthened so that the researcher observes not only the true treatment status but also $\mathbf{1}[x > c]$, which indicates which side of the threshold each observation lies based on the unobserved X. In the absence of the information required in Assumption $A4$ the procedure proposed here will not be feasible. In that case, the approaches of Davezies and Le Barbanchon (2017) and Pei and Shen (2017) may provide potential solutions to the measurement error problem under alternative assumptions regarding the auxiliary data available or the measurement error distributions, respectively.

Assumption A5 is central to our approach, and requires that the $k \leq J$ uncentered moments of the measurement error distribution conditional on the observed mismeasured running variable are identified based on the information in the auxiliary data for each group. This assumption allows these moments to depend on the observed running variable and to differ for each group. Hence, it complements the existing literature on measurement error by permitting dependence in the true and mismeasured running variables, and measurement error that is not identically distributed across groups. This encompasses a large number of empirical applications, as exemplified in Sections 4 and 5. In the very low birth weight example, different hospitals have measurements of varying precision, while in the geographic case regions may be of different size and have different population densities relative to the border.

Intuitively, Assumption A6 states that the measurement error does not provide additional information about the conditional mean of the outcome for both treatment regimes.

2.3 Identification and Estimation

Researchers are faced with two potential identification problems when implementing RD designs in the presence of measurement error. The first source, common to all RD designs, is the potential for local misspecification of the conditional mean function for the outcome. The second is the measurement error itself, which can distort the estimates of the jump at the cutoff. Crucially, the introduction of measurement error renders infeasible the usual local nonparametric approach to deal with the original misspecification problem.

With measurement error a local approach based on shrinking bandwidths ($h \to 0$) around the cutoff is ineffective since the researcher can only observe $\tilde{x}$ and would not be able to guarantee that the true value of the running variable falls within a neighborhood of the cutoff. In other words, even observations that seem close enough to the threshold for treatment might in reality be far away and be a poor comparison to observations just on the other side of the cutoff.

Identification of τ can be obtained through the combination of a polynomial approximation of the conditional mean of the outcome and information about the group specific measurement error distributions.

Theorem 2.1. *Let assumptions A1-A6 hold. Then τ can be identified even if x is not observed.*

See Appendix C for the proof of Theorem 2.1.

To illustrate the problem and the proposed correction, consider the simple case where $\mu_g^{(k)}(\tilde{x}) = \mu_g^{(k)}$ and the (local) quadratic approximations for $f_t(x_{ig})$ for $t = 0, 1$ with parameters $b_{p,t}$ are used:

$$\begin{aligned} f_t(x_{ig}) &= b_{0,t} + b_{1,t}(x_{ig}) + b_{2,t}x_{ig}^2 \\ &= b_{0,t} + b_{1,t}(\tilde{x}_{ig} + e_{ig}) + b_{2,t}(\tilde{x}_{ig} + e_{ig})^2 \\ &= b_{0,t} + b_{1,t}\left(\tilde{x}_{ig} + e_{ig}\right) + b_{2,t}\left[\tilde{x}_{ig}^2 + 2e_{ig}\tilde{x}_{ig} + e_{ig}^2\right] \end{aligned}$$

Note that, conditional on the group affiliation, we have

$$\begin{aligned} E[Y|\tilde{x}, D = t, G] &= E\left[f_t(x_{ig})|\tilde{x}, G\right] \\ &= \left(b_{0,t} + b_{1,t}\mu_g^{(1)} + b_{2,t}\mu_g^{(2)}\right) + \left(b_{1,t} + 2\mu_g^{(1)}\right)\tilde{x}_{ig} + b_{2,t}\tilde{x}_{ig}^2 \qquad (2.2) \\ &= b_{0,t} + b_{1,t}\left(\tilde{x}_{ig} + \mu_g^{(1)}\right) + b_{2,t}\left(\tilde{x}_{ig}^2 + 2\mu_g^{(1)}\tilde{x}_{ig} + \mu_g^{(2)}\right) \qquad (2.3) \end{aligned}$$

Equation (2.2) highlights the problems with using the mismeasured running variable. Specifically, when regressing the observed Y on the mismeasured $\tilde{x}$, the mean outcome for each group

evaluated at $\tilde{x}_{ig} = 0$ will be $b_{0,t} + b_{1,t}\mu_g^{(1)} + b_{2,t}\mu_g^{(2)}$, rather than $b_{0,t}$.[3] Following from equation (A.7), the RD estimate will integrate (2.2) over the distribution of groups on each side of the cutoff giving biased estimates of the intercepts $\widetilde{b}_{0,t} = b_{0,t} + b_{1,t}\overline{\mu_t^{(1)}} + b_{2,t}\overline{\mu_t^{(2)}}$ where $\overline{\mu_t^{(k)}} = E[\mu_{g,t}^{(k)}] = \sum_g P(G=g)\mu_{g,t}^{(k)}$ are the expected values of measurement error moments across the groups over the support of $\tilde{X}$ used in estimation on each side of the cutoff . Taking the difference gives an estimate of the treatment effect $\tau^* = b_{0,1} - b_{0,0} + bias = \tau + bias$: where

$$\begin{aligned} bias &= \left(b_{1,1}\overline{\mu_1^{(1)}} + b_{2,1}\overline{\mu_1^{(2)}}\right) - \left(b_{1,0}\overline{\mu_0^{(1)}} + b_{2,0}\overline{\mu_0^{(2)}}\right) \\ &= \left(b_{1,1}\overline{\mu_1^{(1)}} + b_{2,1}\overline{\mu_1^{(2)}}\right) - \left(b_{1,0}\overline{\mu_0^{(1)}} + b_{2,0}\overline{\mu_0^{(2)}}\right) + \left(b_{1,0}\overline{\mu_1^{(1)}} - b_{1,0}\overline{\mu_1^{(1)}} + b_{2,0}\overline{\mu_1^{(2)}} - b_{2,0}\overline{\mu_1^{(2)}}\right) \\ &= b_{1,0}\left(\overline{\mu_1^{(1)}} - \overline{\mu_0^{(1)}}\right) + \overline{\mu_1^{(1)}}\left(b_{1,1} - b_{1,0}\right) + b_{2,0}\left(\overline{\mu_1^{(2)}} - \overline{\mu_0^{(2)}}\right) + \overline{\mu_1^{(2)}}\left(b_{2,1} - b_{2,0}\right) \end{aligned} \tag{2.4}$$

Equation (2.4) emphasizes the two potential sources of bias from the group-specific measurement error — from differences across the cutoff in either the distribution of groups or the true conditional expectation function. First, if there is a discontinuous change in the distribution of measurement error group types at the cutoff, as in both of our empirical applications, then $\left(\overline{\mu_1^{(1)}} - \overline{\mu_0^{(1)}}\right) \neq 0$ and $\left(\overline{\mu_1^{(2)}} - \overline{\mu_0^{(2)}}\right) \neq 0$, which will introduce bias terms that are proportional to the parameters of the true conditional expectation. Intuitively, the bias from a change in group types at the cutoff is larger when the slope is steeper ($b_{1,0}$ larger) or there is more curvature ($b_{2,0}$ larger) as this magnifies the importance of differences in the measurement error. That is, if the conditional expectations where completely flat with respect to x, then mismeasuring the running variable would not affect our ability to estimate the mean on either side. Even if the distribution of group types is continuous at the cutoff, so that $\left(\overline{\mu_1^{(1)}} - \overline{\mu_0^{(1)}}\right) = 0$ and $\left(\overline{\mu_1^{(2)}} - \overline{\mu_0^{(2)}}\right) = 0$, there is an additional bias if there are differences in the conditional expectation functions so that $(b_{1,1} - b_{1,0}) \neq 0$ and $(b_{2,1} - b_{2,0}) \neq 0$. Intuitively the same measurement error can generate different biases on each side due to how it interacts with changes in the true underlying function.

Equation (2.3) is helpful for understanding the intuition behind our correction procedure. It provides a representation of the mean of the outcome conditional on the variable of interest, the unobserved x, in terms of the observed $\tilde{x}$ and the first J moments of the group-specific measurement error distribution. In other words, one can interpret equation 2.3 as saying that the coefficients of interest can be recovered by fitting y on the "corrected" regressors $x_1^* = \tilde{x}_{ig} + \mu_g^{(1)}$, $x_2^* = \tilde{x}_{ig}^2 + 2\mu_g^{(1)}\tilde{x}_{ig} + \mu_g^{(2)}$ and a constant. In general for any J, the vector of the mismeasured

[3] Under a single homogeneous type of measurement error and known $\mu_g^{(p)}$ Dong (2015) shows that you could correct the estimate for the intercept after the estimation.

running variable $\tilde{X}' = [1, \tilde{x}, \tilde{x}^2, \ldots, \tilde{x}^J]$, is replaced by the vector of the "corrected" running variable of same dimensions $X^{*\prime} = [1, x_1^*, x_2^*, \ldots, x_J^*]$, where $x_j^* = \sum_{k=0}^{j} \binom{j}{k} \mu_g^{(j-k)} \tilde{x}^k$. The uncentered moments for the measurement error can be replaced by consistent estimates.

Take a simplified geographic RD example where $J = 1$. If an individual lives in a county in which the geographic centroid is 25 miles from the state border, but the average individual in that county resides 40 miles from the border, then the mismeasured running variable is $\tilde{x}_{ig} = 25$ and the first uncentered moment of the measurement error distribution for residents of the county is $\mu_g^{(1)} = 15$ — reflecting the fact that the centroid measure is on average wrong by 15 miles. Rather than controlling for distance to the border by the centroid measure, our correction simplifies to the intuitively appealing approach of controlling for the mean distance from the border for residents of that county i.e. our corrected regressor is $x^* = \tilde{x}_{ig} + \mu_g^{(1)} = 25 + 15 = 40$. Also note that the centroid measure would only be appropriate if the population distribution is symmetric around the centroid — a condition that will obviously not hold in practice across all counties. Equation (2.3) extends this example to the case of $J = 2$, and our procedure can be extended to an arbitrary J — which for the geographic case corresponds to controlling for the higher order moments of the underlying population distribution relative to the border, rather than higher powers of the (mismeasured) centroid distance.

More generally, our strategy estimates a regression of order J for treated and untreated observations as described below:

$$\hat{\tau} = \hat{\beta}_+ - \hat{\beta}_- \tag{2.5}$$

$$(\hat{\beta}_+, \hat{\beta}_+^{(1)}, \ldots, \hat{\beta}_+^{(J)})' = argmin_{b_0, b_1, \ldots, b_J} \sum_{i=1}^{N_p} (Y_i - b_0 - b_1 \hat{x}_{1,i}^* - \cdots - b_J \hat{x}_{J,i}^*)^2 \tag{2.6}$$

$$(\hat{\beta}_-, \hat{\beta}_-^{(1)}, \ldots, \hat{\beta}_-^{(J)})' = argmin_{b_0, b_1, \ldots, b_J} \sum_{i=1}^{N_p} (Y_i - b_0 - b_1 \hat{x}_{1,i}^* - \cdots - b_J \hat{x}_{J,i}^*)^2. \tag{2.7}$$

Where N_p is the size of our primary sample on which we observe the outcome and mismeasured running variable. Also, let $\hat{x}_j^* = \sum_{k=0}^{j} \binom{j}{k} \hat{\mu}_g^{(j-k)}(\tilde{x}) \tilde{x}^k$ and $\hat{\mu}_g^{(j)}(\tilde{x})$ be a consistent estimator of $\mu_g^{(j)}(\tilde{x})$.

Since $E(e^k | \tilde{x}, G = g) = \mu_g^{(k)}(\tilde{x})$, several estimators are available in the literature. It is natural to rely on a local estimation methods. In practice, a local polynomial estimator (usually linear) is the preferred choice due to its better control of the leading bias term typical of local

estimation, and good boundary properties (Fan and Gijbels, 1996). Specifically, one could use:

$$\hat{\mu}_g^{(j)}(\tilde{x}) = \arg\min_{\beta_0} \min_{\beta_1} \sum_{i=1}^{N_{a,g}} (e_{i,g}^j - \beta_0 - (\tilde{x}_{i,g} - \tilde{x})\beta_1)^2 K_h(\tilde{x}) \tag{2.8}$$

Where the $N_{a,g}$ is the size of the auxiliary sample for group g on which we observe the measurement error, h_g the group-specific smoothing parameter, and $K_h(\tilde{x}) = K\left(\frac{\tilde{x}_i - \tilde{x}}{h_g}\right)$ is a bounded kernel with usual properties. Furthermore, the local polynomial estimator's properties and bandwidth choice has been extensively studied in recent papers (Calonico, Cattaneo, and Farrell, 2018, 2019b), corroborating the small sample performance improvements of the approach.[4] Therefore, we use a local polynomial estimator to calculate $\hat{\mu}_g^{(j)}(\tilde{x})$ in the simulations in Section 3, and recommend that choice to applied researchers.

2.4 Large Sample Properties

To perform inference about the parameters of interest, we propose a novel studentized test which incorporates the uncertainty associated with estimation for each group of the several moments of the measurement error conditional distributions, $\mu_g^{(j)}(\tilde{x})$, using auxiliary information. Define N_p and $N_{a,g}$ to be the sample sizes for the primary dataset and auxiliary dataset for group g, respectively. Additionally, let $\lambda_g = lim_{N_p \to \infty} \frac{N_p}{N_{a,g} h_g}$ for all g, which essentially requires both N_p and $N_{a,g}h_g$ to go to infinity at the same rate so that the asymptotic approximation captures the effect of estimating $\hat{\mu}_g$. This is a group-level analogue of the asymptotic conditions in Davezies and Le Barbanchon (2017), and would not be needed if identification was not based in auxiliary data (Pei and Shen, 2017; Dong, 2015). Also define e' as the conformable row vector of zeros except for the first entry equal to one.

Theorem 2.2. *Let assumptions A1-A6 hold, and λ_g be a finite constant for every group g. Define, $X_i^{*\prime} = [1, x_{i,1}^*, \ldots, x_{i,J}^*]$, $\mu_g'(\tilde{x}) = [1, \mu_g^{(1)}(\tilde{x}), \ldots, \mu_g^{(J)}(\tilde{x})]$, $B_+' = [\beta_+, \beta_+^{(1)}, \ldots, \beta_+^{(J)}]$ and equivalently for B_-, and $\Gamma_i = L_{J+1} \circ Q$ is the Hadamard product of the lower diagonal Pascal matrix, L_{J+1}, and the matrix Q_i, where $Q_{(b,c)} = \tilde{x}_i^{c-b}$. Also, $\varepsilon = Y - E[Y|\tilde{x}, D, G]$. Finally, assume that $h_g \to 0$, $N_{a,g}h_g \to \infty$ as the sample sizes increase for all g, and that a CLT holds for the vector of measurement error moment estimators using the auxiliary data such that $(N_{a,g}h_g)^{\frac{1}{2}}(\hat{\mu}_g(\tilde{x}) - \mu_g(\tilde{x})) \to N(0, V_g(\tilde{x}))$ for all relevant points of the support of $\tilde{X}$. As*

[4] Implementation for these local polynomial estimators, including several bandwidth selectors, is readily available in commonly used statistical software through the package *nprobust* (Calonico, Cattaneo, and Farrell, 2019a).

$N_p \to \infty$, *then*

$$\sqrt{N_p}(\hat{\tau} - \tau) \to^d N(0, \Omega) \tag{2.9}$$

where,

$$\Omega = e' \left[\Omega_+ + \Omega_-\right] e \tag{2.10}$$

$$\Omega_+ = A^{-1} E\left[X^{*\prime} \varepsilon_i \varepsilon_i' X^*\right] A^{-1} + A^{-1} \left[\sum_{g=1}^{G} \lambda_g F_{+,g}' V_g(\tilde{x}) F_{+,g}\right] A^{-1} \tag{2.11}$$

$$F_{+,g} = E\left[\left(x_{lg}^{*\prime} \otimes B_+' \Gamma_{lg}\right)\right] \tag{2.12}$$

$$A = E\left[X^{*\prime} X^*\right] \tag{2.13}$$

and similarly for Ω_- *on the other side of the cutoff.*

See Appendix C for the proof of Theorem 2.2.

Theorem 2.2 provides the asymptotic approximation to the distribution of $\hat{\tau}$, which we can use for hypothesis testing and creating valid CIs. The asymptotic variance approximation, Ω, incorporates the uncertainty introduced due to the estimation of the parameters μ_g using auxiliary datasets/information for each group-specific measurement error, and can be estimated by a direct plug-in estimator, $\hat{\Omega}$. The explicit adjustment in the variance formula takes into account the amount of information available for estimation of μ_g in each auxiliary dataset in order to obtain CIs with correct coverage when testing hypotheses. The results are related to the auxiliary data/two-sample results in the literature (Chen, Hong, and Tamer, 2005; Lee and Sepanski, 1995) and extend their conclusions to the case in which heterogeneous measurement error is present for groups in the context of RD designs.

2.5 Procedure Summary

We summarize the proposed procedure in the following steps:

Step 1. Determine the appropriate measurement error group definitions for the context.

Step 2. Estimate $\mu_g^{(k)}(\tilde{x})$ for each group in the auxiliary dataset. This could be done by local linear fit on the relevant support of $\tilde{x}$ as described above (Fan and Gijbels, 1996).[5]

Step 3. Match the $\hat{\mu}_g^{(k)}(\tilde{x}_i)$ from the auxiliary data to the appropriate values of $\tilde{x}_i$ observed in the primary dataset.

[5]However, as discussed in our empirical applications, the appropriate way to estimate these moments may differ depending on the nature of the auxiliary data used.

STEP 4. Generate the "corrected" running variable $\hat{x}^*_{j,i} = \sum_{k=0}^{j} \binom{j}{k} \hat{\mu}_g^{(j-k)}(\tilde{x}_i)\tilde{x}_i^k$ for $j = 0, 1, ..., J$ for all observations in the primary dataset.

STEP 5. Estimate the RD treatment effect by the following steps:

1. Choose the polynomial order J_t for $t = 0, 1$ using a cross-validation procedure by estimating equations (2.6) and (2.7) for different potential values of J. We suggest a procedure that rules out "specifications that can be rejected(...)when tested against a strictly more flexible specification" (Lee and Lemieux, 2010, p. 285) using the AIC to compare choices of J:
   (a) Iterate through Steps 2-4 and 5.1 by comparing the AIC for $J = 1$ and $J = 2$. If the AIC is lower for $J = 2$, then compare this to $J = 3$ — progressively adding higher order terms until the AIC does not suggest an improvement.
2. Estimate equation (2.5) using the chosen J from the previous step.

STEP 6. Conduct appropriate inference on $\hat{\tau}$ by obtaining the corrected standard error as the square root of the first element of $\hat{\Omega}$ and calculate the usual CIs based on the asymptotic normality results in Theorem 2.2.

In Step 5, other procedures in line with the suggestion from Lee and Lemieux (2010) are also appropriate. If concerned about misspecification due to the choice of polynomial order, the researcher could consider obtaining honest CIs as described in Section 2.6.

In order to implement this procedure, a researcher needs information about the measurement error distribution moments for each group. As described in Assumption A5, if this information is being acquired from auxiliary data this requires it to represent the same population as the main data, or at least have the same measurement error distribution. Note, however, that it is not necessary that the auxiliary data match specific observations, nor does it need to be nested within the main sample. Furthermore, the auxiliary data needs to contain information about X and $\tilde{X}$ for each group, but could omit the outcome Y. In fact, the auxiliary datasets could be different for each group. These requirements are similar to assumption 3 in Davezies and Le Barbanchon (2017). While these are important assumptions, this highlights the flexibility and transparency of the proposed procedure to accommodate different structures of measurement error by the researcher.

2.6 Honest Confidence Intervals

If the chosen polynomial order,"J," does not fully capture the relevant features of the conditional mean of the outcome, $f_t(\cdot)$, this may lead to misspecification bias in the treatment effect estimate

and potentially invalid inference based on the approximations discussed in the previous section. We address this concern by proposing honest CIs that build upon the insights and techniques developed by Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b). Honest CIs cover the true parameter at the nominal level uniformly over all possible functions (in a family) for the conditional mean of the outcome, $f_t(\cdot)$, thereby providing valid inference even in the presence of potential misspecification bias.

Here we focus on the class of functions that place bounds on the derivatives of f_t, with M denoting the smoothness of the functions being considered, which is chosen by the researcher. Armstrong and Kolesár (2018b) provide a rule of thumb procedure for M based on the largest derivative of a flexibly estimated global model of the outcome which we discuss in detail in Appendix B.1. These honest CIs are built by considering the non-random *worst-case bias* of the estimator $\hat{\tau}$ for functions that respect the bounds on its derivatives.[6] Intuitively, the researcher considers what would be the worst possible misspecification bias that could arise in the estimation of τ if we use a polynomial of order "J" to approximate f_t, under the assumption that the $(J+1)-th$ derivative of the true function is bounded by a constant and adjusts the CIs used for inference accordingly. Our approach approximates $f_t(\cdot)$ with a polynomial of order J that can be recovered from the mismeasured observed data following the procedures described in Section 2.3. It then obtains an approximation to the worst-case bias that can be used to generate the honest CIs. In this setting, misspecification errors can be rewritten as:

$$f_t(x) = \sum_{j=0}^{J} x_j \beta_j + R_t(x), \text{where } |R_t(x)| \leq \bar{R}_t(x), \tag{2.14}$$

where $\bar{R}_t(x)$ is a bound on the misspecification error that is chosen by the researcher.

By the same arguments used in Theorem 2.1 to map from x to the corrected x^*, we can then sum across the individuals in each group to show that

$$N_g^{-1} \sum_{i=1}^{N_g} f(x_i) = N_g^{-1} \sum_{i=1}^{N_g} \sum_{j=0}^{J} x_{j,i}^* \beta_j + N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) + o_p(1). \tag{2.15}$$

Hence we can obtain the worst-case bias by focusing on $N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) = R_{t,g}$. The bounds for $|R_{t,g}|$ will be directly obtained from the conditions and class function adopted for the original $f_t(x)$. Note that these bounds are defined in terms of the original conditional means of the outcome, considering the model with no measurement error on the running variable. This

[6] For more details, see Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b).

property is attractive since the researcher is likely to have better guidance from economic theory in terms of the true running variable.

We focus on the Taylor and Hölder classes of functions in our proposed honest CIs since, as indicated by Armstrong and Kolesár (2018) and Armstrong and Kolesár (2018b), those are natural function families to consider in the RDD setting. For example, assume f_t is such that it can be approximated by polynomial of order "J" and is a member of the following Taylor class of functions:

$$\left\{ f : \left| f(x) - \sum_{j=0}^{J} f^{(j)}(0)\frac{x_j}{j!} \right| \leq \frac{M}{p!}|x^{J+1}|, \text{for all } x \in \mathcal{X} \right\} \tag{2.16}$$

Intuitively, if we were to approximate the true conditional mean of the outcome by a Taylor polynomial of order J, the misspecification errors at each value of X will be bounded by the $(J+1)$-th derivative of the function, which we assume is at most equal to "M" chosen by the researcher based on his beliefs about the smoothness of the conditional mean. Then, by similar arguments to the ones used in Equation 2.15, we can rewrite the bounds on misspecification in terms of the observed transformed data.

$$|R_t(x)| \leq \frac{M}{p!}|x^{J+1}| \tag{2.17}$$

$$|R_{t,g}| \leq \frac{M}{p!}N_g^{-1}\sum_{i=1}^{N_g}|x^*_{J+1,i}| + o_p(1) \tag{2.18}$$

Hence, we can rely on the results in Armstrong and Kolesár (2018) and Armstrong and Kolesár (2018b) to obtain honest CIs in the presence of measurement error as described above. In particular, the worst case bias could be written as

$$\overline{bias}_M(\hat{\tau}) = \frac{M}{p!}\sum_{i=1}^{n}|w^n_+(x^*_i) + w^n_-(x^*_i)||x^*_{J+1,i}|. \tag{2.19}$$

where $w^n_+(x^*_i)$ and $w^n_-(x^*_i)$ are the first elements of the usual $(X^{*\prime}X^*)^{-1}X^{*\prime}$ separately for each treatment group — reflecting the relative weight put on each observation when estimating the mean outcome at $x = 0$. One can think of Equation (2.19) as the worst possible distortion that would be introduced to the estimate of τ by leaving the $(J+1)$-th order of the polynomial in the error term if the $(J+1)$-th derivative of $f_t(x)$ was as steep as possible, for a choice of

smoothness (M). The honest CIs can then be obtained as

$$\hat{\tau} \pm cv_\alpha \left(\frac{\widehat{\overline{bias}}_M(\hat{\tau})}{\hat{se}_n} \right) \cdot \hat{se}_n. \tag{2.20}$$

where $\widehat{\overline{bias}}_M(\hat{\tau})$ replaces X^* with its feasible counterparts, $\hat{se}_n = \hat{\Omega}^{\frac{1}{2}}$, the variance matrix Ω and its estimators are those described in Theorem 2.2, and $cv_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ distribution. Note that the critical values are from a normal distribution re-centered around the worst case bias, and will differ from the usual critical values (e.g., ± 1.96) used when we assume the correct specification. Nicely enough, we can apply this approach directly to the estimators discussed in Section 2.3 regardless of how we choose the polynomial order "J." Additional details of this honest CI approach, including a discussion of choice of M and implementation, are available in Appendix B.

3 Simulation Evidence

This section presents simulation evidence on the performance of the RD measurement error correction.

3.1 Data Generating Process

For ease of comparison with the previous literature, we focus on a DGP similar to Calonico, Cattaneo, and Titiunik (2014) except that the running variable may be measured with error. Throughout we still define X to be the true running variable and $\tilde{X}$ to be the mismeasured running variable observed by the researcher. The simulated data are generated as follows:

$$\begin{aligned} \tilde{X}_i &\sim U(-1,1), \\ X_i &= \tilde{X}_i + \epsilon_i, \\ Y_i &= m_j(X_i) + v_i, \end{aligned}$$

where $v_i \sim i.i.d.\ N(0, 0.1295^2)$. The expectation of the outcome conditional on X, is given by:

$$m(x) = \begin{cases} 0.52 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } D = 1 \\ 0.48 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise,} \end{cases}$$

The model is based on a modified fifth-order polynomial fitted to the data in Lee (2008) in

his analysis of incumbency effects on electoral races to the U.S. House of Representatives. We then introduce seven different types of measurement error, ϵ_i, with the following distributions:

1. $U(0, a)$: "rounding down" with uniform distribution,
2. $U(-a, 0)$: "rounding up" with uniform distribution,
3. $U(-a, a)$: "rounding to midpoint" with uniform distribution,
4. $N(\mu, \sigma^2)$ truncated by $(0, a)$: "rounding down" with truncated normal distribution,
5. $N(\mu, \sigma^2)$ truncated by $(-a, 0)$: "rounding up" with truncated normal distribution,
6. $N(\mu, \sigma^2)$ truncated by $(-a, a)$: "rounding to midpoint" with truncated normal distribution,
7. No measurement error.

where $a = 0.1$, $\mu = 0.05$, $\sigma = 0.05$. Each observation i is randomly assigned to one of these seven groups from which the measurement error will be drawn, and these "group" assignments are observed by the researcher.

We run eight separate simulation scenarios that differ across three dimensions. We consider two treatment determination mechanisms – one in which treatment is determined by the true, unobserved running variable ($D = \mathbf{1}[x < 0]$) and one in which the treatment is determined by the mismeasured, observed running variable ($D = \mathbf{1}[\tilde{x} < 0]$). The former is consistent with the settings typically considered in the RD measurement error literature, while the latter reflects relevant empirical cases, including the very low birth weight analysis covered in our empirical applications. We also differentiate between cases in which the polynomial order is correctly specified ("Known J"), in this case $J = 5$, and the case in which we must select J in some way ("Select J"). The first case is useful for isolating the performance of our correction in the absence of misspecification of the polynomial order, while the latter encompasses any bias from using a potentially misspecified polynomial. Importantly, this second case matches the situation faced in many applications. In the current simulations, we pick J by using the Akaike Information Criteria (AIC). Finally, we consider different sample sizes with a "small" sample of 500 primary and 1000 auxiliary observations and "large" sample with 5000 primary and 10000 auxiliary observations.

Across our eight simulation scenarios we present the empirical bias and coverage rates based on a 5% nominal size test for the null hypothesis that τ equals its true value for a set of estimators and inference procedures. For the Known J case, we consider the following three estimates:

1. "Naive" - Mismeasured running variable without a measurement error correction.

2. "Corrected" - Mismeasured running variable with a measurement error correction using estimated error moments and adjusted standard errors.

3. "No Error" - Infeasible case using the true running variable.

In the Select J case which introduces the possibility of mispecifying the polynomial order, we also present the empirical coverage from applying the honest confidence intervals from Section 2.6 to our adjusted estimator, which we label as "Adjusted with HCI".

Panel A of Table 3.1 displays the results for the Known J case. Starting with the small sample size with treatment determined by the true unobserved running variable, the adjusted estimator reduces the bias relative to the naive estimator by over half, from 0.1165 to 0.0441, while significantly improving the coverage from 13.80% to 90.40%, achieving coverages close to the infeasible "no measurement error in the running variable" case. Similarly, we see a greater reduction in the bias from -0.172 to -0.0069 in the case with treatment determined by the mismeasured running variable, while the empirical coverage of the adjusted estimator closely matches that for the "no error" estimate in this case as well. Note that the bias for the naive estimator stays stable when increasing the sample size, while the relative performance of the adjusted estimates further improves with smaller bias and coverage rates very close to 95% in both treatment determination cases.

In Panel B, we no longer assume the researcher knows the correct order of the polynomial and include a data driven choice of J into the estimation procedure as described above. For small samples, the potential misspecification can be seen by the slightly larger bias and worse coverage for the "no error" and adjusted estimators. Nonetheless, both still outperform the naive estimator. Importantly, when we apply the honest CI to the adjusted estimator we see the coverage improve to 98.50% and 95.55%, respectively, when treatment is determined based on the correctly and mismeasured running variables. Finally, increasing the primary and auxiliary sample sizes results in similar bias for the naive estimate while the adjusted and "no error" estimators improve in both bias and coverage by being able to better approximate the true conditional mean function. Once again, the honest CI perform well in this setting and provide coverage rates close to 95%.

Appendix D presents an additional simulation comparing our approach to that from Davezies and Le Barbanchon (2017) in the two-sided fuzzy setting in the presence of measurement error that changes at the cutoff.

Table 3.1: Simulation Evidence

Panel A: Known J					
	Treatment:	$D = \mathbf{1}[x < 0]$		$D = \mathbf{1}[\tilde{x} < 0]$	
Sample Size	*Estimator*	*Bias*	*Coverage*	*Bias*	*Coverage*
Small	Naive	0.1165	0.1380	-0.1272	0.7335
	Adjusted	0.0441	0.9040	-0.0069	0.9280
	No Error	0.0001	0.9325	-0.0009	0.9290
Large	Naive	0.1155	0.0000	-0.1239	0.0160
	Adjusted	0.0075	0.9440	-0.0009	0.9410
	No Error	0.0004	0.9510	-0.0001	0.9410
Panel B: Select J					
	Treatment:	$D = \mathbf{1}[x < 0]$		$D = \mathbf{1}[\tilde{x} < 0]$	
Sample Size	*Estimator*	*Bias*	*Coverage*	*Bias*	*Coverage*
Small	Naive	0.1061	0.1170	-0.0985	0.6725
	Adjusted	0.0608	0.6915	-0.0163	0.8515
	Adjusted with HCI		0.9850		0.9555
	No Error	0.0244	0.7450	-0.0029	0.8910
Large	Naive	0.1090	0.0000	-0.1178	0.0575
	Adjusted	0.0201	0.8625	-0.0040	0.8805
	Adjusted with HCI		0.9820		0.9670
	No Error	-0.0001	0.9135	-0.0002	0.9505

Simulation results based on 2000 replications with a small sample of 500 primary and 1000 auxiliary observations and a large sample of 5000 primary and 10000 auxiliary observations.

4 Application I: Very Low Birth Weight

Here we apply our approach to the case studied by ADKW looking at the effect on infant mortality of additional care received by newborns classified as Very Low Birth Weight (VLBW). They take advantage of the fact that VLBW is classified based on having a measured birth weight of strictly less than 1500 grams. This setup lends itself to estimating the effect of these additional resources and services by RD where the measured birth weight is the running variable and treatment is switched on when passing 1500g from above.

ADKW focus on a window of measured birth weights from 1415g-1585g and estimate the treatment effect controlling for a linear function in birth weight on each side of the cutoff. Doing so, they estimate fairly large effects of additional care. Their baseline estimates with no controls suggest a 0.95 percentage point decline in the one-year mortality rate from crossing the 1500g threshold and receiving additional care, a fairly large effect given a mean mortality rate of 5.53 percent for the untreated just above the cutoff.[7]

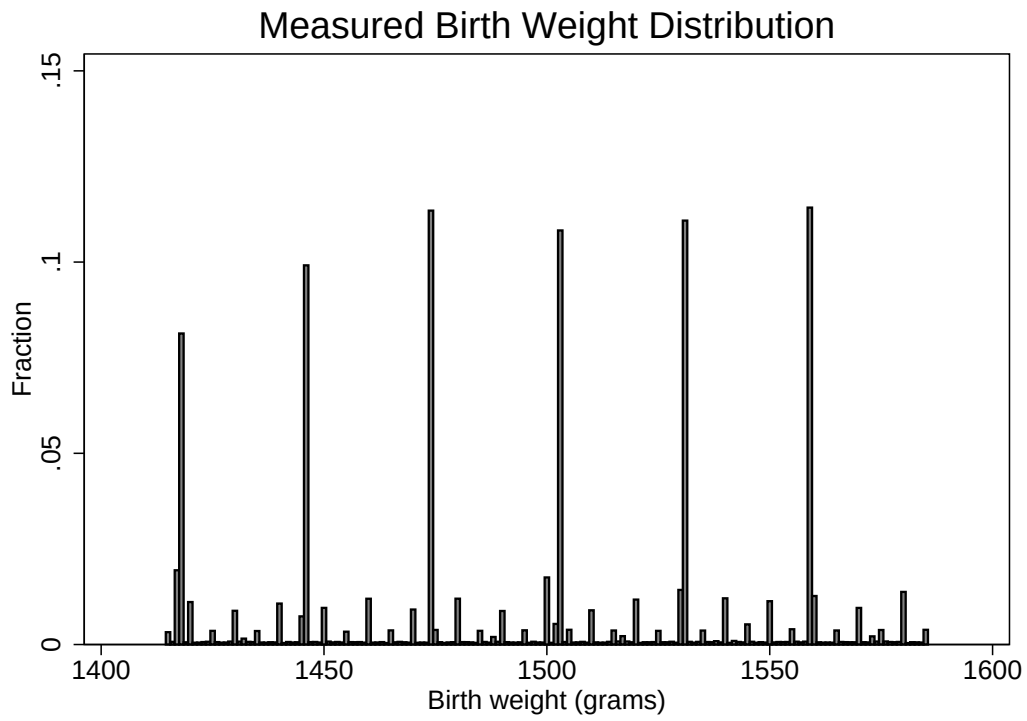
BGLW suggest caution in interpreting these results by noting that the observed distribution of birth weights shows large "heaps" at ounce multiples and at multiples of 100g as well as smaller heaps at other points (multiples of 50g, 25g, etc.). BGLW focus on the fact that some of the heaped measures tend to have higher mortality rates than neighboring unheaped measures. In particular, they emphasize that the observations measured at 1500g have "substantially higher mortality rates than surrounding observations on *either* side of the VLBW threshold." BGLW view this as evidence of potential non-random sorting into a 1500g birth weight measure and propose a simple sensitivity check called the Donut RD. They test the robustness of the estimated treatment effect to dropping the heaped observations very near the cutoff, creating a "donut hole" with no data around the cutoff. They start by dropping the 1500g observations and then progressively increase the size of the donut hole until they exclude observations with measured birth weights between 1497g-1503g. Importantly, 1503g corresponds to a large heap at 53 ounces. BGLW find that the estimated treatment effect falls substantially when omitting observations near the cutoff.

[7] ADKW's main results include additional control variables, but here we focus on the RD results without controls. See Frölich and Huber (2018) for a discussion of RD with and without covariates.

4.1 Data

The main data on birth weight and infant mortality are drawn from the National Center for Health Statistics linked birth/infant death files.[8] The data are discussed in detail in ADKW and BGLW. Briefly, the data include information from birth certificates for all births in the US between 1983-1991 and 1995-2002 and are linked to death certificates for infants up to one year after birth. For the main analysis sample used here this yields 202,078 separate births with an overall one year mortality rate of 5.8 percent. The histogram in Figure 4.1 shows the distribution of measured birth weights for the main sample used with the largest heaps occurring at the six ounce multiples within the 1415g-1585g window used by ADKW and BGLW.

Figure 4.1



Source: National Center For Health Statistics Linked Birth-Infant Death Files. N=202,078.

4.2 Measurement Error Correction

The analysis in BGLW highlights the potential importance of heaping in running variables. Here, we extend their analysis by addressing the underlying measurement problem that leads

[8] Raw data files available at `https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm`. We thank Alan Barreca for providing the data files from BGLW.

to heaping.[9] Heaping at ounce and gram multiples is likely due to rounding errors. Specifically, BGLW note that the scales used by hospitals to weigh newborns differ in their precision and there may be a human tendency to round numbers when recording the birth weight. Ideally, we would like to know the precision of the scales used to measure the birth weight of each baby in order to determine the measurement error groups, since the resolution of the scale determines the range of latent birth weights for a given observed weight. However, such information on the scales used is not available, so instead we approximate this group structure under plausible scenarios given the data available. Here we explore the importance of the differential rounding error by using our measurement error correction, and assume that measured birth weights at ounce, 100g, 50g, 25g, 10g, or 5g multiples reflect true birth weights that were rounded to that nearest multiple. All other observed measures— those not at one of the multiples— are assumed to be correctly measured. For instance, it is assumed that the true birth weight for those measured at 1500g will range from 1450g to 1549g and were simply rounded to the nearest 100g multiple. Similarly, those measured at 1503g (53oz) had true birth weights between half an ounce above and below (from 1489g-1517g).

This sort of differential rounding leads to an interesting pattern of potential measurement errors. In Figure 4.2, we plot the observed birth weight measure on the vertical axis and the range of potential true birth weights on the horizontal distinguishing between observed measures that receive treatment (observed measure less than 1500g) and those that do not. First note that among those with a measured birth weight just to the right of the cutoff, many may have true birth weights well below the cutoff. This provides a potential explanation for why the mortality rate at 1500g is noticeably higher— namely, these children do not receive the additional care, but many will have similar birth weights and associated unobservable factors to children at much lower birth weights who do receive additional treatment. Also note that this "misclassification" only occurs for untreated units as none of the true weight ranges for treated units cross the threshold. Finally, note that the 1500g measure exhibits the largest potential measurement error range in the described rounding error scenario.

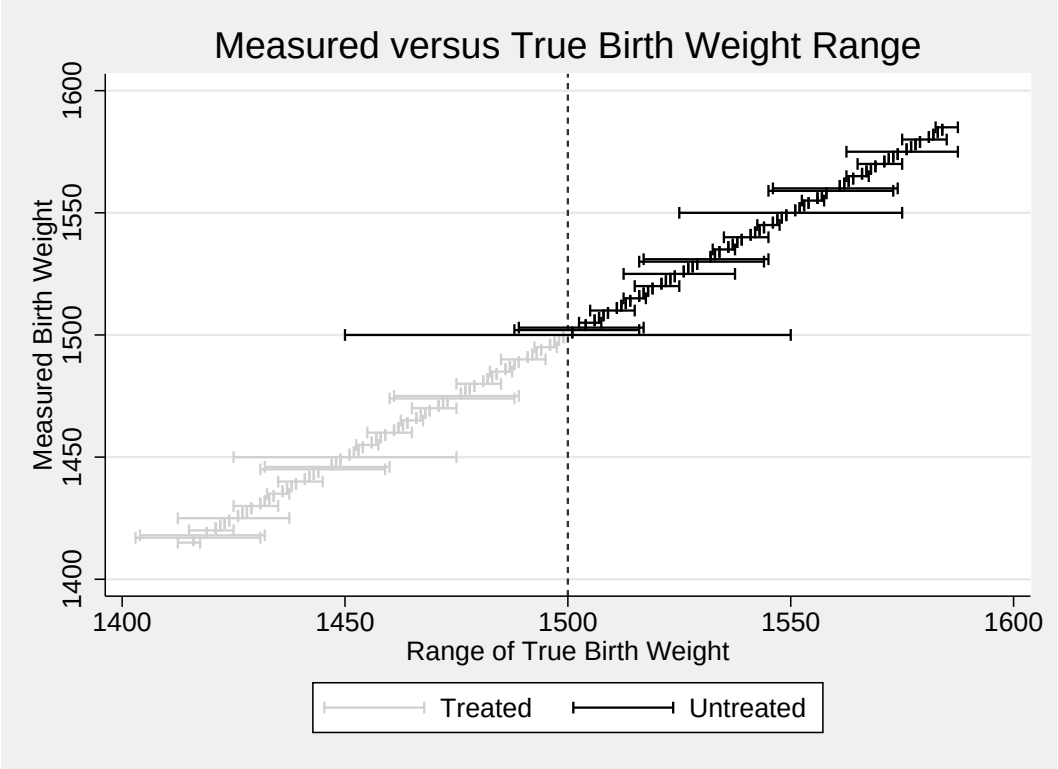
Figure 4.2 also suggests that this setting fits well with our correction procedure: there are groups of observations that face different measurement error distributions and these groups are — in part — related to the measured birth weight. Our method requires approximating the true birth weight distributions within each measurement group. In the context of our procedure, unheaped data serves as auxiliary data to estimate the distribution of the measurement error.

[9]Note that while our correction accounts for potential discontinuities in measurement error near the cutoff, it does not address potential endogeneity in hospital measuring systems, similarly to the previous literature.

In particular, for our baseline estimates, we use all births with observed birth weights between 1000g-2000g that are not at one of the heaped values and obtain a kernel density estimate of the distribution for the unheaped observations — denoted $\hat{f}(x)$.[10] The estimated density is used to calculate the moments of the measurement error distribution in each measurement group by estimating the density weighted mean of $(x - \tilde{x})^k$ over the support of true x associated with each observed measure $\tilde{x}_g$ denoted $[\underline{x}_g, \overline{x}_g]$ as follows:

$$\hat{\mu}_g^k(\tilde{x}_g) = \frac{\sum_{x=\underline{x}_g}^{\overline{x}_g} (x - \tilde{x}_g)^k \hat{f}(x)}{\sum_{x=\underline{x}_g}^{\overline{x}_g} \hat{f}(x)} \tag{4.1}$$

Figure 4.2



Ranges refer to the range of potential true birth weights for a given measured birth weight.

In Table 4.1, we present the corrected and uncorrected estimates for different samples. As suggested by Lee and Lemieux (2010), we use a cross validation procedure, the Akaike Information Criteria (AIC), to choose the polynomial's order on either side of the VLBW cutoff. To do so, we first estimate the first eight uncentered moments of the measurment error distributions

[10] Nearly identical results were obtained including all births— both heaped and unheaped— when estimating the density while using a wide bandwidth in order to smooth out the heaps. This suggests that in terms of estimating the true birth weight distribution our choice to focus on unheaped measures only does not lead to a problematic selection issue.

and generate the eight corresponding corrected running variable terms — allowing us to test the fit of up to an eighth order polynomial. We then estimate each side separately using our corrected running variables with $J = 1$ and add higher order terms until the AIC no longer suggests an improvement in fit. In this case, the procedure gave $J = 4$ for the untreated and $J = 1$ for the treated. Starting in the first row, using the same sample as BGLW, we see a large difference between the uncorrected and corrected estimates in Columns (a) and (b), respectively. The uncorrected estimates suggest a 3.1 percentage point drop in the mortality rate when receiving additional care, while the corrected estimate is a drop of only 0.67 percentage points. Importantly, when we calculate an honest CI using the procedure in Appendix B for our corrected estimate, the upper bound of the 95% honest CI still does not overlap zero. This suggests that the estimated reduction in mortality is robust to misspecification of the conditional mean function.

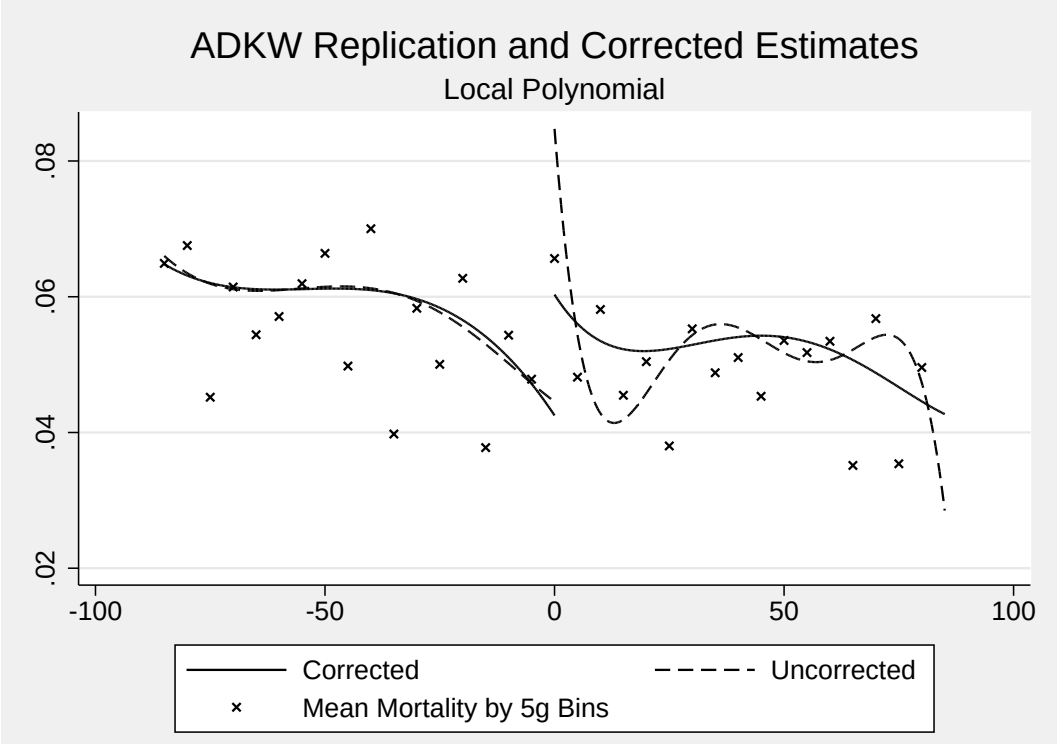
Table 4.1

RD VLBW Estimates: Naive and Corrected		
Estimator:	*Naive*	*Corrected*
	(a)	*(b)*
Panel A: Main Estimates		
BGLW Sample	-0.0311	-0.0067
$N_p = 202,078$	(0.0037)	(0.0025)
Honest CI		[-0.0121, -0.0013]
Panel B: Measurement Error Group Sensitivity		
$\widetilde{X}$ *by Education*	-0.0311	-0.0067
$N_p = 202,078$	(0.0037)	(0.0031)
Panel C: Donut RD Sensitivity		
(1) Omitting 1500g	-0.0139	-0.0067
$N_p = 198,530$	(0.0043)	(0.0025)
(2) Omitting 1499-1501g	-0.0149	-0.0069
$N_p = 198,334$	(0.0043)	(0.0025)
(3) Omitting 1498-1502g	-0.0156	-0.0069
$N_p = 197,135$	(0.0045)	(0.0025)
(4) Omitting 1497-1503g	-0.0038	0.0102
$N_p = 175,108$	(0.0147)	(0.0110)

Source: National Center For Health Statistics Linked Birth-Infant Death Files. $\widetilde{X}$ Groups: $N_a = 27,846; \min(N_{a,g}) = 10,145; \max(N_{a,g}) = 12,279$. $\widetilde{X}$ by Mother's Education Groups: $N_a = 27,846; \min(N_{a,g}) = 1,033; \max(N_{a,g}) = 5,067$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Standard errors in parentheses with adjusted standard errors for the corrected estimates. Honest CIs in square brackets.

To provide some intuition for the difference in the two estimates, Figure 4.3 depicts the estimated functions on either side of the cutoff along with mean mortality rates within five gram bins of the observed birth weight measure. First, we see that the two approaches yield similar estimates of the conditional mean function to the left of the cutoff. However, we see very different fitted functions to the right of the cutoff. Intuitively, the uncorrected function gets very steep near the cutoff because it treats the observations at a measured weight of 1500g that have a higher mean mortality rate as being precisely measured at 1500g and tries to fit that point. In contrast, our correction recognizes that many of those with a measured birth weight of 1500g may have a true birth weight away from 1500g and the regression function above the cutoff is not influenced as much by these observations.

Figure 4.3



Source: National Center For Health Statistics Linked Birth-Infant Death Files. $N_p = 202,078; N_a = 27,846; \min(N_{a,g}) = 10,145; \max(N_{a,g}) = 12,279$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Lines represent local polynomial regressions with order chosen using AICc.

An additional issue raised by BGLW is that the measurement technology available may differ by hospitals that serve women with different backgrounds. In particular, hospitals in higher poverty areas may have less precise scales (more likely to have a rounded birth weight). If the true birth weight distribution differs across different maternal backgrounds— for example,

more mass at lower birth weights for disadvantaged mothers— this could lead to differences in the measurement error distributions for babies at the same observed measure. To address this possibility, we allow the measurement error distributions to differ by mother's education level (less than high school, high school, some college, college and above, and missing education data). Specifically, we simply redefine our measurement error groups to be based on the observed measure and mother's education. We then re-estimate the birth weight density for each education level to generate a new set of corrected moments. Panel B of Table 4.1 displays the results using the mother's education specific measurement groups. The results are very similar to those in Panel A, with an estimated treatment effect of 0.0067 with adjusted standard errors that are slightly larger due to the fact that moment correction terms are estimated using smaller samples.

We also revisit the donut RD from BGLW here, holding the order of the polynomial fixed as in BGLW. In Panel C, Row (2) of Table 4.1, we see that the corrected estimate is unaffected by dropping the 1500g heap, while the uncorrected falls by over half as in BGLW. The fact that the corrected estimate is robust to dropping the 1500g heap is encouraging that our correction is helping to control for the underlying measurement problem that led to the heap. Intuitively, since the uncorrected estimator treats every observation measured at 1500g as precisely measured and this group has a relatively high mortality rate, omitting these observations removes a large mass with a high mortality rate from a single point right at the cutoff. Instead, the corrected procedure accounts for the fact that most observations measured at 1500g actually have true birth weights above or below the threshold. Not only has our correction "smoothed out" the heap at 1500g, it also accounts for the fact that the true birth weights for observations measured at 1503g includes the 1500g cutoff. Therefore, dropping observations at the observed heap of 1500g is similar to randomly dropping some of the observations from a range of true X while keeping the cutoff in the support of the data — an adjustment that would not lead to large differences in estimates in any case.

While dropping observations with measured birth weights up to 3g away from the cutoff alters the corrected estimate in Row (4), the corrected estimate otherwise appears remarkably stable across the different size donut holes. In contrast, uncorrected estimates are quite sensitive to the different size donut holes used. Importantly, 1503g corresponds to 53oz and ounce measures seem to be the most common type in the data. As ADKW note in their reply to BGLW, this actually removes about 20 percent of the data to the right of the cutoff while barely dropping any to the left (Almond et al., 2011). This is because the closest ounce measure from below is at

1474g. In justifying their approach, BGLW note that dropping those within 3g and at the cutoff represents an "incremental" difference in birth weights since the "implied gap in birth weights between the observations to the left and right of the cutoff is roughly equivalent in weight to seven paper clips (7g)." However, when viewed from the perspective that those measured at ounce multiples are rounded to the nearest ounce, this implies that the gap between most true birth weights when dropping 1503g is actually between 29g-85g since the largest ounce measure below the cutoff is 1474g with a true range from 1460-1488g and the first ounce measure above 1503g is at 1531g with a true range of 1517-1545g. In particular, now most the data on the untreated side are for babies with much higher birth weights who have much lower mortality rates regardless of treatment. This suggests some caution considering the basic RD identification argument when dropping the 1503g heap as the babies on either side of the threshold may no longer be comparable along unobservables.

5 Application II: UI Benefit Effects using Geographic RD

In this section, we apply our correction procedure to the problem of estimating the effect of Unemployment Insurance (UI) extensions on unemployment during the Great Recession using a GeoRD. During the Great Recession, the duration of UI benefits was extended from 26 weeks to as many as 99 weeks. The realized benefit duration varied at the state level and was determined by state-level labor market aggregates passing pre-specified trigger levels.[11] In theory, such extensions may lead to increased unemployment through reduced job search effort by workers and a contraction of vacancies by firms. The main econometric challenge in estimating the effect on unemployment is to isolate the differences due to the policy from the differences due to the factors driving adoption of the policy.

Hagedorn et al. (2015) and Dieterle, Bartalotti, and Brummet (Forthcoming) both study this case in detail, attempting to exploit differences in UI extensions at state boundaries in estimation.[12] Here, the goal is to compare the preferred RD estimates using the measurement error correction proposed in Section 2 to those using a mismeasured, centroid-based, distance to the state borders. During the recession, there were many instances in which neighboring states faced different UI regimes due to the fact that the extensions were triggered by state

[11] See Hagedorn et al. (2015) and Rothstein (2011) for a more detailed discussion of the institutional details of Unemployment Insurance benefit extensions.

[12] Dieterle, Bartalotti, and Brummet (Forthcoming) implements the measurement error correction procedure as proposed in this paper for the whole U.S. Note that the focus of that paper is very different and, in particular, it does not discuss how the centroid and corrected estimators differ — the key focus of our current empirical investigation.

level aggregate unemployment. To focus our discussion on the correction procedure, we will consider one such case: the Minnesota-North Dakota boundary in the second quarter of 2010. The average available UI benefit duration over the entire quarter in Minnesota was 62 weeks while it was only 43 weeks in North Dakota.

5.1 Data

We use county-level data on the unemployment rate from the Bureau of Labor Statistics' (BLS) Local Area Unemployment Statistics (LAUS), and the duration of UI benefits provided by US Department of Labor.[13] Our sample includes all counties located in either state for which the MN-ND boundary is the closest state boundary.

5.2 Measurement Error

The main issue with implementing the RD strategy in this case is that geographic location is reported at the county level, but the underlying running variable is a continuous measure of distance to the border. Researchers often calculate the distance to the border based on the geographic center of the county. This geographic centroid based distance measure is the mismeasured running variable in this context. To implement the measurement error correction in this example, we require auxiliary data on the measurement error distribution (in this case, the distribution of the true distance to the border). To obtain estimates of this distribution, we use data on the geographic location of counties and the within-county population distribution relative to a state boundary from the TIGER geographic shapefiles containing population counts by census block from the 2010 Census. The geographic information gives precise location of census block, county, and state borders. For the centroid-based distance we can therefore calculate the distance from the geographic center of a county to the state border. We can also calculate the distance from the center of each census block to the state boundary. Since census blocks are typically very small, we can use this to approximate a continuous measure of the population weighted distance to the border needed for our measurement error correction.

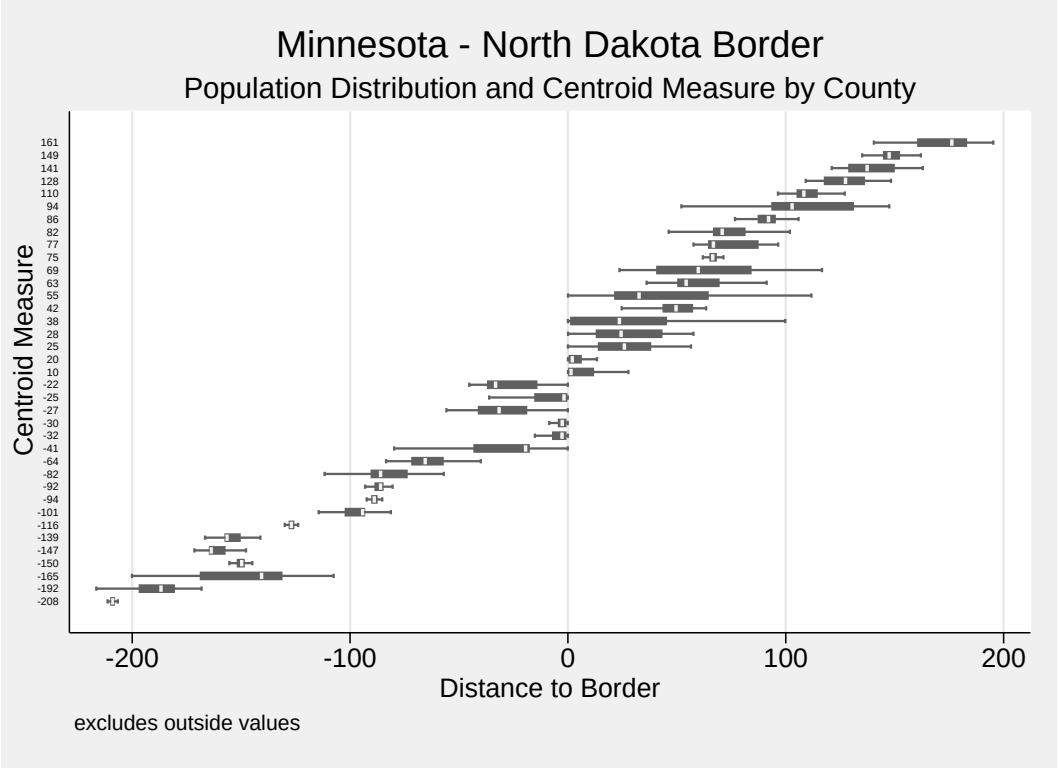
To calculate the population moments and the centroid-based distance from the TIGER shapefiles, we use the nearstat package in Stata (Jeanty, 2010). Since particular areas within a county may have a different nearest neighbor, we determine the modal nearest state boundary among the census blocks in a county and then calculate distances and moments based on the

[13] See `http://ows.doleta.gov/unemploy/trigger/` and `http://ows.doleta.gov/unemploy/euc_trigger/`. Here, we use the county-level unemployment rate as given. See Dieterle, Bartalotti, and Brummet (Forthcoming) for a discussion of potential issues with using an aggregate outcome measure in this setting.

modal neighbor. This gives us the population by distance from the border in each county which we use to calculate the corrected running variable as the population weighted moments of the distance measure — i.e. $x^*_{k,g}$ is the population weighted mean of x^k_{ig} in the TIGER shapefiles.

Figure 5.1 depicts box plots of the population distribution within each county in our sample. Here, the population distributions are directly linked to the measurement error distributions. We have ordered the plots by the centroid measure, starting at the county farthest from the border in North Dakota 208 km away up to the farthest in Minnesota at 161 km away. Several features of the measurement error are worth highlighting. First, the measurement error distributions do not cross the cutoff, so the treatment is identically defined by the true and mismeasured running variables. In many cases, when comparing two counties the one that is measured to be closer by the centroid based measure actually has most of the population mass farther away. For many of the counties the population distributions are far from symmetric and, importantly, they vary substantially across each group (county) at the border. Together this suggests that the group-specific measurement error correction may be particularly important in this setting.

Figure 5.1



Source: US Census TIGER Geographic Shapefiles. N=93,530 census blocks

5.3 Results

We estimate the effect of the difference in available UI duration at the Minnesota-North Dakota border in the second quarter of 2010 on log unemployment by GeoRD using both the uncorrected centroid based measure and our moments based correction. We first generate the first eight corrected running variable terms using the auxiliary shapefile data and then choose J for each side using our corrected estimator with different choices of J on each side comparing the small sample version of the Akaike Information Criteria (AICc) due to the relatively small sample size on both sides of the border.

Table 5.1 presents the uncorrected and corrected estimates for the ATE at the boundary. The uncorrected estimate is large and negative, but imprecise. The point estimate for the uncorrected case would suggest a 25 percent reduction in unemployment from the 19 extra weeks of UI available in Minnesota. The corrected estimate is much smaller in magnitude— nearly zero— and more precisely estimated. However, the honest confidence interval shown in brackets indicate much less precision when accounting for potential misspecification in this setting. The fact that the honest CI is wide here is consistent with the simulation evidence suggesting larger potential misspecification bias in smaller samples. The lack of an estimated effect when using our correction is consistent with the evidence of UI policy spillovers discussed in Dieterle, Bartalotti, and Brummet (Forthcoming).

Table 5.1

Geographic RD: Corrected and Uncorrected Estimates

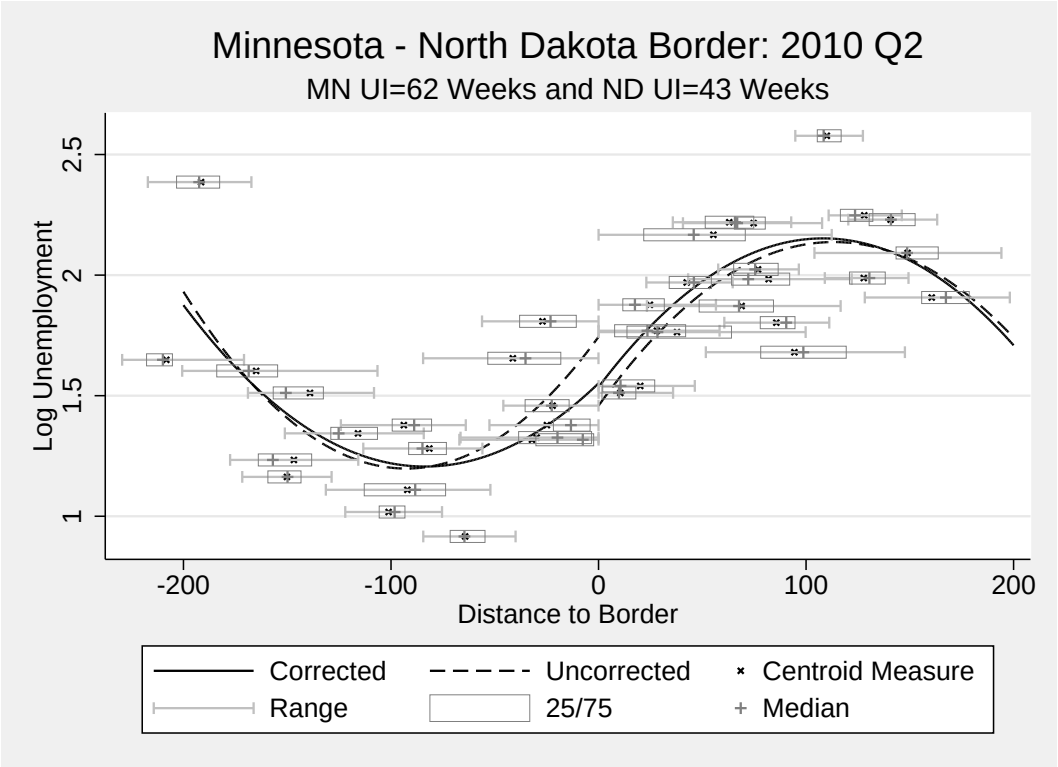
(1)	(2)
Naive	*Corrected*
-0.2855	-0.0155
(0.1943)	(0.1548)
	[-1.0936, 1.0626]

Source: LAUS and TIGER Geographic Shapefiles. $N_p = 38; N_a = 93,530; \min(N_{a,g}) = 845; \max(N_{a,g}) = 7,921$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Standard errors in parentheses with adjusted standard errors for the corrected estimates. Honest CIs in square brackets.

To provide more intuition for the correction procedure, Figure 5.2 depicts the corrected and uncorrected estimated polynomials along with the centroid measures. We also overlay the range, twenty-fifth to seventy-fifth percentile range, and the median of the population distribution to provide some sense of how the population distribution differs from the centroid measure. On the North Dakota side of the border (distance less than zero), we see the uncorrected polynomial is influenced by a few counties that have a centroid distance roughly 10-40km away from the

border, but have households living right up to the border. The corrected estimates take into account the fact that the population distribution is skewed toward the state border for many of these counties lowering the estimated intercept on the North Dakota side. On the Minnesota side, the effect is the opposite, raising the estimated intercept. Combined, this reduces the estimated treatment effect.

Figure 5.2



Source: LAUS and TIGER Geographic Shapefiles. $N_p = 38; N_a = 93,530; \min(N_{a,g}) = 845; \max(N_{a,g}) = 7,921$. N_p is the primary sample size, N_a is the total auxiliary sample size, and $N_{a,g}$ is the auxiliary sample size for group g. Lines represent local polynomial regressions with order chosen using AICc.

6 Conclusion

RD designs have become increasingly popular in empirical studies, but researchers often face situations where there are several types of group-specific measurement errors in the forcing variable. In order to accommodate these situations, we propose a new procedure that utilizes auxiliary information to correct for the bias induced by group-specific measurement error. We develop a valid estimator of the RD treatment effect and derive asymptotic variance formulas that take into account both the variability introduced by the measurement error correction

and the use of multiple data sets in estimation. This method complements previous work on measurement error in RD designs by allowing more flexible forms of the measurement error, including measurement error that is potentially non-classical and discontinuous at the cutoff. Furthermore, the approach is effective regardless of whether treatment is assigned based on the "true" or mismeasured running variable. Finally, we also provide honest CIs that allow for valid inference even under certain forms of misspecification in the conditional mean function.

Simulation evidence supports the theoretical results proposed on the paper and its superior performance relative to "naive" estimators. In two empirical illustrations, we demonstrate that correcting for measurement error can provide a new empirical perspective on the data.

7 Acknowledgments

We would like to thank Alan Barreca for kindly sharing the data used in one of our applications. We also thank the two anonymous referees for comments that led to substantial improvement to this paper. Finally, we are indebted to Yang He for his invaluable research assistance, Tim Armstrong, Cristine Pinto, Sergio Firpo, Samuele Centorrino, Jeff Wooldridge, Christian Hansen, Alfonso Flores-Lagunes and participants at presentations at Michigan State University, 2017 Midwest Econometrics Group Meeting, 2017 Brazilian Econometric Society Meeting, 2018 North American Summer Meeting of the Econometrics Society, 2018 European Summer Meeting of the Econometric Society and 2018 CMStatistics for valuable comments.

References

Almond, Douglas, Joseph J Doyle, Amanda E Kowalski, and Heidi Williams. 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns." *Quarterly Journal of Economics* 125 (2):591–634.

Almond, Douglas, Joseph J. Doyle, Jr., Amanda E. Kowalski, and Heidi Williams. 2011. "The Role of Hospital Heterogeneity in Measuring Marginal Returns to Medical Care: A Reply to Barreca, Guldi, Lindo, and Waddell*." *The Quarterly Journal of Economics* 126 (4):2125–2131. URL +http://dx.doi.org/10.1093/qje/qjr037.

Armstrong, Timothy and Michal Kolesár. 2018b. "Simple and honest confidence intervals in nonparametric regression." Cowles Foundation Discussion Paper.

Armstrong, Timothy B. and Michal Kolesár. 2018. "Optimal Inference in a Class of Regression Models." *Econometrica* 86 (2):655–683. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14434.

Barreca, Alan I, Melanie Guldi, Jason M Lindo, and Glen R Waddell. 2011. "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification." *Quarterly Journal of Economics* 126 (4):2117–2123.

Barreca, Alan I, Jason M Lindo, and Glen R Waddell. 2016. "Heaping-Induced Bias in Regression-Discontinuity Designs." *Economic Inquiry* 54 (1):268–293.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4):588–638.

Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *Quarterly Journal of Economics* 114 (2):577–599.

Calonico, Sebastian, Matias Cattaneo, and Max Farrell. 2019a. "nprobust: Nonparametric Kernel-Based Estimation and Robust Bias-Corrected Inference." *Journal of Statistical Software, Articles* 91 (8).

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2018. "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference." *Journal of the American Statistical Association* 113 (522):767–779. URL https://doi.org/10.1080/01621459.2017.1285776.

———. 2019b. "Coverage error optimal confidence intervals for local polynomial regression." *arXiv:1808.01398* .

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6):2295–2326.

Chen, Xiaohong, Han Hong, and Elie Tamer. 2005. "Measurement Error Models with Auxiliary Data." *Review of Economic Studies* 72 (2):343–366.

Davezies, Laurent and Thomas Le Barbanchon. 2017. "Regression Discontinuity Design with Continuous Measurement Error in the Running Variable." *Journal of Econometrics* 200 (2):260–281.

Dell, Melissa. 2010. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78 (6):1863–1903.

Dieterle, Steven, Otávio Bartalotti, and Quentin Brummet. Forthcoming. "Revisiting the Effects of Unemployment Insurance Extensions on Unemployment: A Measurement-Error-Corrected Regression Discontinuity Approach." *American Economic Journal: Economic Policy* .

Dong, Hao. 2017. "Sharp Regression-discontinuity Design with a Mismeasured Running Variable." Working Paper.

Dong, Yingying. 2015. "Regression Discontinuity Applications with Rounding Errors in the Running Variable." *Journal of Applied Econometrics* 30 (3):422–446.

Eugster, Beatrix, Rafael Lalive, Andreas Steinhauer, and Josef Zweimüller. 2011. "The Demand for Social Insurance: Does Culture Matter?" *Economic Journal* 121 (556):F413–F448.

Falk, Oliver, Robert Gold, and Stephan Heblich. 2014. "E-lections: Voting Behavior and the Internet." *American Economic Review* 104 (7):2238–2265.

Fan, Jianqing and Irene Gijbels. 1996. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, vol. 66. CRC Press.

Frölich, Markus and Martin Huber. 2018. "Including covariates in the regression discontinuity design." *Journal of Business & Economic Statistics* (just-accepted).

Gibbons, Stephen, Stephen Machin, and Olmo Silva. 2013. "Valuing School Quality Using Boundary Discontinuities." *Journal of Urban Economics* 75:15–28.

Hagedorn, Marcus, Fatih Karahan, Iourii Manovskii, and Kurt Mitman. 2015. "Unemployment Benefits and Unemployment in the Great Recession: The Role of Macro Effects." Federal Reserve Bank of New York Staff Reports No. 646.

Hausman, Jerry A., Whitney K. Newey, Hidehiko Ichimura, and James L. Powell. 1991. "Identification and estimation of polynomial errors-in-variables models." *Journal of Econometrics* 50 (3):273 – 295. URL `http://www.sciencedirect.com/science/article/pii/0304407691900226`.

Jeanty, P.W. 2010. "nearstat: Stata Module to Calculate Distances, Generate Distance-based Variables, and Export Distance to Text Files." URL `http://ideas.repec.org/c/boc/bocode/s457110.html`.

Keele, Luke and Rocío Titiunik. 2014. "Natural Experiments Based on Geography." *Political Science Research and Methods* :1–31.

Koop, Gary M. 2003. *Bayesian econometrics*. John Wiley & Sons Inc.

Lalive, Rafael. 2008. "How Do Extended Benefits Affect Unemployment Duration? A Regression Discontinuity Approach." *Journal of Econometrics* 142 (2):785–806.

Lavy, Victor. 2006. "From Forced Busing to Free Choice in Public Schools: Quasi-Experimental Evidence of Individual and General Effects." NBER Working Paper No. 11969.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2):675–697.

Lee, David S. and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142 (2):655–674.

Lee, David S. and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2):281–355.

Lee, Lung-fei and Jungsywan H Sepanski. 1995. "Estimation of Linear and Nonlinear Errors-in-variables Models Using Validation Data." *Journal of the American Statistical Association* 90 (429):130–140.

Pei, Zhuan and Yi Shen. 2017. "The Devil is in the Tails: Regression Discontinuity Design with Measurement Error in the Assignment Variable." In *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics)*, vol. 38, edited by Matias D. Cattaneo and Juan Carlos Escanciano. 455–502.

Rothstein, Jesse. 2011. "Unemployment Insurance and Job Search in the Great Recession." *Brookings Papers on Economic Activity* 43 (2):143–213.

Yu, Ping. 2012. "Identification of Treatment Effects in Regression Discontinuity Designs with Measurement Error." Working Paper, University of Auckland.

A Naive Estimators' Identification with Heterogeneous Measurement Error

A.1 Intuition

To gain some intuition about the problems introduced by mismeasurement, consider the special case in which the conditional distribution of the measurement error is continuous. In that case, a researcher that ignores the measurement error and implements standard RD techniques will estimate:

$$\tau^* = \lim_{a \downarrow 0} E\left[Y|\tilde{X} = c + a\right] - \lim_{a \uparrow 0} E\left[Y|\tilde{X} = c + a\right] \tag{A.1}$$

$$= \int_{-\infty}^{\infty} (y_1(c+e)p_1(c+e) + y_0(c+e)p_0(c+e)) f_{e|\tilde{X}}(e|\tilde{X} = c^+)de$$

$$- \int_{-\infty}^{\infty} (y_1(c+e)p_1(c+e) + y_0(c+e)p_0(c+e)) f_{e|\tilde{X}}(e|\tilde{X} = c^-)de \tag{A.2}$$

where $p_0(X)$ and $p_1(X)$ are the probabilities of receiving and not receiving treatment conditional on the unobserved X and $f_{Z|W}(\cdot|W = c^+)$ and $f_{Z|W}(\cdot|W = c^-)$ denote a conditional density of the variable Z evaluated as W approaches c from above and below, respectively. By looking at the neighborhood of $\tilde{X} = c$ we are effectively analyzing the (weighted) average of the potential outcomes over the values of X for which $\tilde{X} = c$. This quantity in Equation (A.1) will take different forms based on whether treatment is assigned on the observed or unobserved running variable.

If treatment is sharply assigned based on the true unobserved running variable, standard RD techniques will estimate:

$$\tau^* = \int_{X \geq c} y_1(x)(f_{x|\tilde{X}}(x|\tilde{X} = c^+) - f_{x|\tilde{X}}(x|\tilde{X} = c^-))dx$$

$$+ \int_{X < c} y_0(x)(f_{x|\tilde{X}}(x|\tilde{X} = c^+) - f_{x|\tilde{X}}(x|\tilde{X} = c^-))dx \tag{A.3}$$

which equals zero in the absence of a discontinuity in $f_{x|\tilde{X}}(x|\tilde{X} = c)$ (or equivalently $f_{e|\tilde{X}}(e|\tilde{X} = c)$). This is an example of the loss identification induced by the presence of continuous measurement error in the running variable described in Pei and Shen (2017) and Davezies and Le Barbanchon (2017) in which the measurement error smooths out the conditional expectation of the outcome close to the observed "cutoff" in $\tilde{X}$. Intuitively, this occurs because the measurement error induces the misclassification of treatment to some observations.

If instead treatment is determined by the mismeasured running variable and is therefore observed, a researcher that ignores the measurement error and implements standard RD techniques will estimate:

$$\tau^* = \lim_{a \downarrow 0} E\left[Y|\tilde{X} = c + a\right] - \lim_{a \uparrow 0} E\left[Y|\tilde{X} = c + a\right] \tag{A.4}$$

$$= \lim_{a \downarrow 0} \int_{-\infty}^{\infty} y_1(x) f_{x|\tilde{X}}(x|\tilde{X} = c + a)dx - \lim_{a \uparrow 0} \int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X}}(x|\tilde{X} = c + a)dx$$

$$= \int_{-\infty}^{\infty} y_1(c + e) f_{e|\tilde{X}}(e|\tilde{X} = c^+)de - \int_{-\infty}^{\infty} y_0(c + e) f_{e|\tilde{X}}(e|\tilde{X} = c^-)de. \tag{A.5}$$

If the measurement error distribution conditional on X is continuous at the policy cutoff, then:

$$\tau^* = \int_{-\infty}^{\infty} (y_1(x) - y_0(x)) \frac{f_{\tilde{X}|X}(c|x)}{f_{\tilde{X}}(c)} dF_X(x) \tag{A.6}$$

Hence, instead of estimating the ATE at the cutoff, the researcher recovers a weighted average treatment effect for the population in the support of X for which $X + e = \tilde{X} = c$. The weights are directly proportional to the *ex ante* likelihood that an individual's value of $\tilde{X}$ will be close to the threshold. This case is similar to the situation described in Lee (2008) and Lee and Lemieux (2010) where individuals can manipulate the running variable with imperfect control. Our approach will recover $E[Y_1 - Y_0|X = c]$ in this setting as well. Note that while this case includes both the geographic RD and birth weight examples discussed in the main body of the paper, our procedure also applies to the case where treatment is assigned based on the unobserved running variable provided that the researcher observes true treatment status.

A.2 Group-specific Measurement Error Distribution

A central contribution of this paper is to allow for heterogeneity of the measurement error distribution across groups of observations. Let $\tilde{x}_{ig} = x_i - e_{ig}$, where e_{ig} is the measurement error of "type" g. This notation allows each unit to have a measurement error drawn from a separate, group-specific distribution. It also encompasses the case where individuals in the same group share the same observed value of $\tilde{X}$, such that $\tilde{x}_{ig} = \tilde{x}_g$ for all individuals in group g. This is the situation where, for example, residents in a county have their location reported as the county's centroid or birth weights being rounded to nearest 50 grams or ounce multiples.

Once we allow for different measurement error "types," the identification of the ATE is further complicated by the averaging across groups on both sides of the cutoff. Discontinuous changes in the share of groups at the cutoff introduce bias to estimates of the treatment effect.

For example, if all individuals follow the same processes $y_1(x)$ and $y_0(x)$ but suffer from different types of measurement error in the running variable and treatment is assigned based on $\tilde{X}$, then

$$\tau^* = \sum_g P(G=g|\tilde{X}=c^+) \int_{-\infty}^{\infty} y_1(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^+, G=g)dx \tag{A.7}$$
$$- P(G=g|\tilde{X}=c^-) \int_{-\infty}^{\infty} y_0(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^-, G=g)dx.$$

Hence, changes in the share of each group at the cutoff could introduce bias and measurement error correction approaches that ignore the group heterogeneity might fail to identify the intended ATE.

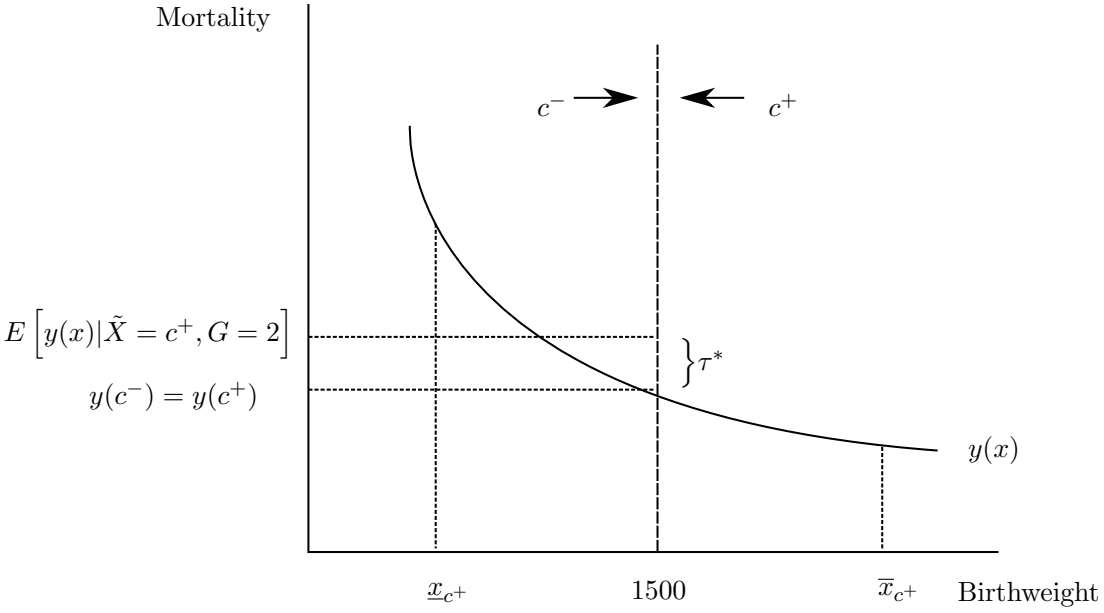
To illustrate the problem, consider the very low birth weight example. For now, assume there are two types of measures— those correctly measured at the individual gram level ($G=1$) and those rounded to a gram multiple ($G=2$), where this could be any of the gram-multiple heaps observed in the data (5g, 10g, 25g, 50g, or 100g). Further assume for exposition that treatment has no effect on outcomes so that $y_0(x) = y_1(x) = y(x)$ and $\tau = 0$. Here, treatment is turned on by passing the threshold from above, so we adjust notation so that c^- denotes treated and c^+ denotes the untreated side. In the limit, the conditional expectation for treated units just below the cutoff will come from the children measured at 1499g— all of which are correctly measured (i.e. $P(G=1|\tilde{X}=c^-)=1$). Meanwhile the untreated units at 1500g will be a mixture of correctly measured and mismeasured units. This implies the following RD estimand:

$$\tau^* = y(c^-) - \left[P(G=1|\tilde{X}=c^+)y(c^+) + P(G=2|\tilde{X}=c^+) \int_{-\infty}^{\infty} y(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^+, G=2)dx \right]$$

If the probability of being a rounded measure close to the cutoff is quite high relative to precise measures, that is if $P(G=2|\tilde{X}=c^+)$ is large, the estimate of the conditional expectation for untreated units will be driven by the mismeasured group. In the very low birth weight data used in Section 4, the observations at exactly 1500g make up 1.75 percent of the overall sample while the adjacent unrounded measure of 1501g only makes up 0.05 percent of the sample, implying $P(G=2|\tilde{X}=c^+) \approx 0.97$. Given evidence of rounding to 100g multiples in some cases, $\int_{-\infty}^{\infty} y(x) f_{x|\tilde{X},G}(x|\tilde{X}=c^+, G=2)dx$ may average the outcome (mortality) over a range of true x between 50 grams below to 50 grams above the cutoff. How much that averaging will impact the estimated outcome for the untreated at the cutoff depends on the shape of $y(x)$. If, for instance, the mortality rate decreases with birth weight, but at a decreasing rate as we approach the natural lower bound of zero ($y'(x) < 0$ and $y''(x) < 0$), as is likely the case in the

low birth weight example, this may lead to a very poor estimate of the intended estimand of $\lim_{a\downarrow 0} E[Y|X = c + a]$. We will likely overestimate the conditional expectation at the cutoff for the untreated units since $\int_{-\infty}^{\infty} y(x) f_{x|\tilde{X},G}(x|\tilde{X} = c^+, G = 2)dx > y_0(c^+)$ by Jensen's Inequality. This is depicted in Figure A.1.[14] Our proposed approach will be able to recover identification of the ATE in these settings.

Figure A.1



B Honest CIs under Measurement Error: Further Details

As discussed in Section 2.6, if the polynomial order chosen,"J," does not fully capture the relevant features of the conditional mean of the outcome, $f_t(\cdot)$, some misspecification bias may arise, potentially invalidating inference based on the approximations discussed in Section 2.4. We address this concern by proposing novel honest CIs that build upon the insights and techniques developed by Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b). Honest CIs cover the true parameter at the nominal level uniformly over the possible parameter space for $\mathcal{F}(M)$ for $f_t(\cdot)$. Here we focus on the class of functions that place bounds on the derivatives of f_t, with M denoting the smoothness of the functions being considered, which is chosen by the researcher. These honest CIs are built by considering the non-random *worst-case bias* of

[14]Let the range of true birth weights associated with a measured weight of c be denoted by $[\underline{x}_c, \overline{x}_c]$. Then, for simplicity, the true birth weight is assumed to be uniformly distributed for units with measured weight equal to 1500g, i.e. $f_{x|\tilde{X},G}(x|\tilde{X} = c^+, G = 2)$ is uniform in $[\underline{x}_{1500}, \overline{x}_{1500}] = [1450, 1550]$.

the estimator $\hat{\tau}$ for functions in $\mathcal{F}(M)$. For more details, see (Armstrong and Kolesár, 2018; Armstrong and Kolesár, 2018b). Intuitively, the researcher considers what would be the worst possible misspecification bias that could arise in the estimation of τ under the assumption that the true conditional mean of the outcome is part of a class of functions (which typically places bounds on higher derivatives of $f_t(\cdot)$) and adjusts the CIs used for inference accordingly.

By applying the insights about identification and inference in the main body of the text we can extend the honest CIs approach in Armstrong and Kolesár (2018) to the measurement-error-corrected RD setting. Intuitively we will approximate $f_t(\cdot)$ by a polynomial of order J which can be recovered from the observed data following the procedures described in Section 2.3 and obtain an approximation to the worst-case bias that can be used to generate the honest CIs.

In this setting, misspecification errors in $f_t(\cdot)$ can be rewritten as:

$$f_t(x) = \sum_{j=0}^{J} x_j \beta_j + R_t(x), |R_t(x)| \leq \bar{R}_t(x) \tag{B.1}$$

$$f_t(x) = \sum_{j=0}^{J} x_j^* \beta_j + R_t^*(x) \tag{B.2}$$

where

$$R_t^*(x) = \sum_{j=0}^{J} \left[\sum_{k=0}^{j} \binom{j}{k} \left(e^{(j-k)} - \mu_g^{(j-k)} \right) \tilde{x}^k \right] \beta_j + R_t(x) \tag{B.3}$$

Note that we can sum across the individuals on each group,

$$N_g^{-1} \sum_{i=1}^{N_g} f(x_i) = N_g^{-1} \sum_{i=1}^{N_g} \sum_{j=0}^{J} x_{j,i}^* \beta_j + N_g^{-1} \sum_{i=1}^{N_g} R_t^*(x_i) \tag{B.4}$$

$$N_g^{-1} \sum_{i=1}^{N_g} R_t^*(x_i) = N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) + o_p(1) \tag{B.5}$$

Hence we can obtain the worst-case bias by focusing on $N_g^{-1} \sum_{i=1}^{N_g} R_t(x_i) = R_{t,g}$. The bounds for $|R_{t,g}|$ will be directly obtained from the conditions and class function adopted for the original $f_t(x)$. It is worth noting that these are defined in terms of the original conditional means of the outcome, considering the model with no measurement error on the running variable. That is positive since the properties and constraints the researcher imposes on the DGP are more natural in that setting for which the researcher has a better intuition and guidance from economic theory.

We focus on the Taylor and Hölder class of functions in our proposed honest CIs since, as

indicated by Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b) those are natural function families to consider in the RDD setting. For concreteness, consider the Taylor class of functions defined as

$$f_t \in \mathcal{F}_J(M) = \left\{ f : \left| f(x) - \sum_{j=0}^{J} f^{(j)}(0)\frac{x_j}{j!} \right| \leq \frac{M}{p!}|x^{J+1}|, \text{for all } x \in \mathcal{X} \right\}. \tag{B.6}$$

Then, by similar arguments we can rewrite the bounds on misspecification in terms of the observed transformed data.

$$|R_t(x)| \leq \frac{M}{p!}|x^{J+1}| \tag{B.7}$$

$$|R_{t,g}| \leq \frac{M}{p!}N_g^{-1}\sum_{i=1}^{N_g}|x_i^{J+1}| \leq \frac{M}{p!}N_g^{-1}\sum_{i=1}^{N_g}|x^*_{J+1,i}| + o_p(1) \tag{B.8}$$

Hence, we can rely on the results in Armstrong and Kolesár (2018); Armstrong and Kolesár (2018b) to obtain honest CIs in the presence of measurement error as described above.

In particular, the estimator proposed in Section 2.3 can be rewritten to match more closely the notation in those papers

$$\hat{\tau} = \sum_{i=1}^{n} w^n(x_i^*)y_i, \tag{B.9}$$

$$w^n(x^*) = w_+^n(x^*) - w_-^n(x^*) \tag{B.10}$$

$$w_+^n(x^*) = e_1' Q_{n,+}^{-1} X^* D \tag{B.11}$$

$$Q_{n,+} = \sum_{i=1}^{n} D_i X_i^{*\prime} X_i^* \tag{B.12}$$

And similarly for $Q_{n,-}$ and weights $w_-^n(x^*)$. Note that when we replace $w^n(x^*)$ by $\hat{w}^n(\hat{x}^*)$ an additional term will be added to the estimator's residuals. This analysis fits within the framework developed by Armstrong and Kolesár (2018) Theorem F.1. with the following adjustments in notation:

$$\hat{L} = \hat{\tau} = \sum_{i=1}^{n} \hat{w}^n(\hat{x}_i^*)y_i \tag{B.13}$$

$$u_i = \hat{x}_i^{*\prime}\left[(x_i^* - \hat{x}_i^*)\, B_+ + \varepsilon_i\right] \tag{B.14}$$

with $s_{n,\mathcal{Q}} = \Omega^{\frac{1}{2}}$, and $\hat{se}_n = \hat{\Omega}^{\frac{1}{2}}$ where the variance matrix Ω and its estimators are those described in Theorem 2.2. Then, coupling the bias bounds derived above with the results in

Armstrong and Kolesár (2018), the largest possible bias of the estimator over the parameter space $\mathcal{F}_J(M)$ is asymptotically given by

$$\overline{bias}_{\mathcal{F}_J(M)}(\hat{L}) = \frac{M}{p!}\sum_{i=1}^{n}|w_+^n(x_i^*) + w_-^n(x_i^*)||x_{J+1,i}^* - |. \tag{B.15}$$

and the honest CIs can be obtained as

$$\hat{L} \pm cv_\alpha\left(\frac{\widehat{\overline{bias}}_{\mathcal{F}_J(M)}(\hat{L})}{\hat{se}_n}\right)\cdot \hat{se}_n. \tag{B.16}$$

where $cv_\alpha(t)$ is the $1-\alpha$ quantile of the absolute value of a $N(t,1)$ distribution, and $\widehat{\overline{bias}}_{\mathcal{F}_J(M)}(\hat{L})$ replaces $w^n(x^*)$ by $\hat{w}^n(\hat{x}^*)$.

B.1 Honest Confidence Interval Implementation

In order to implement the honest CI procedure, we need to estimate the bias term in equation (B.16). To do so we need to first determine the class of functions and the smoothness, M. In our simulations and applications, we follow Armstrong and Kolesár (2018b) in assuming f_t belongs to the Hölder class which assumes smoothness globally:

$$f_t \in \mathcal{F}_J(M) = \left\{f : \left|f^J(x) - f^J(x')\right| \leq M\left|x - x'\right|, \text{for all } x \in \mathcal{X}\right\}. \tag{B.17}$$

Under the Hölder class assumption, it is convenient to rewrite the bias term as in Appendix C.2 of Armstrong and Kolesár (2018b):

$$\overline{bias} = M\int\left|\overline{w}_{J+1}(s)\right| ds \tag{B.18}$$

$$\overline{w}_{J+1}(s) = \frac{1}{nh}\sum_{i=1}^{n}\frac{\tilde{w}^n(x^*)\left(x_i - s\right)^J}{J!}\mathbf{1}(x_i \geq s) \tag{B.19}$$

where

$$\tilde{w}^n(x^*) = w^n(x^*)(nh) \tag{B.20}$$

Where we have used the fact that our polynomial approximation is equivalent to using a rectangular kernel over the whole support of x with the implied bandwidth denoted by h. Importantly, $\overline{w}_{J+1}(s)$ is written in terms of the weights used in estimation (rescaled to sum to the product nh) and the distribution of the true x which we observe in the auxiliary data. Therefore, we can promptly estimate $\overline{w}_{J+1}(s)$ using the weights from our main estimation combined with the

auxiliary data. In our simulations, we first calculate $Q_{n,+}$ and $Q_{n,-}$ from the primary data set as in equation (B.12), but then create our kernel weights by replacing X^* in equation (B.11) with the adjusted X^* from the auxiliary data — effectively determining the weight that each auxiliary observation would have received if it had been in the primary sample. We then rescale the resulting weight to sum to the auxiliary sample size times the implied bandwidth — which is simply the range of x associated with each treatment status. With this estimate of $\tilde{w}^n(x^*)$ in hand, we then calculate $\overline{w}_{J+1}(s)$ for all potential values of s.

In our empirical applications, the estimation of $\overline{w}_{J+1}(s)$ differs slightly due to the nature of the data used. In both cases, the measurement error groups are uniquely defined by a distinct, discrete observed $\tilde{x}$ so that we can apply the exact weights used in our primary sample estimation to the auxiliary sample. For example, in the geographic case all observations in the same county have the same observed centroid measure. We simply need to match the implied weight to the underlying true running variable distribution. To implement this, we use our estimated population densities from the census block level data and rescale the weights to sum to the total of the density at each distance times h, instead of nh, and multiply the expression in equation (B.19) by the estimated density.

In order to calculate the bias term in (B.18), we also need to choose the smoothness of the functions to be considered, M. Armstrong and Kolesár (2018b) propose a rule of thumb choice of M that we follow in our empirical applications. Specifically, they suggest setting M equal to the largest, in absolute value, $J+1$-th derivative from a global $J+3$ order polynomial approximation to the conditional mean function. Importantly, in our setting this requires estimating the $J+3$ approximation using our measurement error correction. To summarize the procedure in our empirical applications — we repeat steps 2-4 and 5.2 from section 2.5 for a $J+3$ order polynomial. We then calculate the $J+1$-th derivative of the estimated polynomial over a fine grid of points in the support of the running variable - and then pick M to be the largest — in absolute value — calculated derivative in the support of X.

In our simulations, we alter the Armstrong and Kolesár (2018b) rule of thumb since preliminary simulations suggested that a direct application of their rule of thumb led to incorrectly large confidence intervals. Specifically, their setting focused on providing honest CI for an estimator that relies on an arbitrarily chosen and fixed local polynomial order — typically local linear — while using the bandwidth as the main tuning parameter to improve the fit. In our setting, since the measurement error makes the bandwidth an inappropriate tool to improve fit, we fix the bandwidth and use the polynomial order as the key tuning parameter. Therefore, when we have

chosen a higher J to better approximate the conditional expectation function globally we run the risk that the $J+3$ approximation will severely overfit the relationship leading to unreliable estimates of the $J+1$-th derivatives.

To overcome this, we reduce the number of additional terms added to the approximation and estimate it so that additional terms receive weight related to the extra information they provide. Specifically, we estimate the parameters of a $J+1$ order polynomial, denoted $\widetilde{\mathbf{b}}_{J+1}$ as a weighted average of our main estimates for the J order polynomial— augmented with a zero for the $J+1$-th term — $\hat{\mathbf{b}}_{J,0}$ and the $J+1$ order estimates $\hat{\mathbf{b}}_{J+1}$ using our correction where the weights are proportional to the relative variances of the adjusted X^* used in each case. This is equivalent to a Bayesian regression using our main estimates as the prior for the first J terms of $\mathbf{b}_{J+1}$ and an uninformative prior on the $J+1$ term — effectively estimating the final term based on the additional information provided by the added higher order term (Koop, 2003). More formally, we set M equal to the maximum $J+1$-th derivative using the following Bayesian regression estimates based on the discussion in Koop (2003):

$$\begin{aligned}
\tilde{\mathbf{b}}_{J+1} &= (V_{J,0} + V_{J+1})^{-1} \left(V_{J,0}\hat{\mathbf{b}}_{J,0} + V_{J+1}\hat{\mathbf{b}}_{J+1} \right) \\
V_{J+1} &= X^{*\prime}_{J+1} X^*_{J+1} \\
V_{J,0} &= \begin{bmatrix} X^{*\prime}_J X^*_J & \mathbf{0}'_J \\ \mathbf{0}_J & 0 \end{bmatrix}
\end{aligned} \tag{B.21}$$

where $\mathbf{0}_J$ is a $1 \times J$ vector of zeros and X^*_J and X^*_{J+1} include the first J and $J+1$ adjusted regressors. In practice, this approach works well in our simulations. In our applications we maintain the more conservative approach based on the Armstrong and Kolesár (2018b) rule of thumb — leaving the optimal choice of M when using a global polynomial approximation for future research.

C Proofs

Proof of Theorem 2.1

Proof. First note that, under Assumption A2, we can rewrite the conditional expectation $f_t(x)$ using a polynomial of order J for $t = 0, 1$ with unknown coefficients b_{jt}.

$$E\left[Y|x, D=t\right] = f_t(x) = \sum_{j=0}^{J} b_{jt} x^j \tag{C.1}$$

Let these coefficients be collected in the column vector $B_t' = [b_{0t}, b_{1t}, \ldots, b_{Jt}]$. Then, since for each observation $x = \tilde{x} + e$,

$$f_t(x) = \sum_{j=0}^{J} b_{jt}(\tilde{x} + e)^j = \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} e^{j-k}\tilde{x}^k \tag{C.2}$$

where $\binom{j}{k}$ is the binomial coefficient $\frac{j!}{k!(j-k!)}$. Then,

$$\begin{aligned}
E\left[Y|\tilde{x}, D=t, G\right] =& E\left[E\left[Y|x, D=t\right]|\tilde{x}, D=t, G\right] = E\left[f_t(x)|\tilde{x}, G\right] && \text{(C.3)}\\
=& E\left[\sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} e^{j-k}\tilde{x}^k|\tilde{x}, G\right] && \text{(C.4)}\\
=& \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} E\left[e^{j-k}|\tilde{x}, G\right]\tilde{x}^k && \text{(C.5)}\\
=& \sum_{k=0}^{J}\sum_{j=k}^{J} \binom{j}{k} b_{jt} \mu_g^{(j-k)}(\tilde{x})\tilde{x}^k && \text{(C.6)}\\
=& \sum_{j=0}^{J} b_{jt}\left[\sum_{k=0}^{j} \binom{j}{k} \mu_g^{(j-k)}(\tilde{x})\tilde{x}^k\right] && \text{(C.7)}
\end{aligned}$$

By Assumption $A5$, the first J uncentered moments of the measurement error distribution for each group, $\mu_g^{(j-k)}(\tilde{x})$, are known or estimable. The last equality is simply rewriting the sum for convenience. Let $x_j^* = \sum_{k=0}^{j} \binom{j}{k}\mu_g^{(j-k)}(\tilde{x})\tilde{x}^k$ and $X^{*\prime} = [1, x_1^*, \ldots, x_J^*]$, then the expectation of the outcome Y conditional on the observed $\tilde{x}$ and treatment status can be written as

$$E\left[Y|\tilde{x}, D=t, G\right] = \sum_{j=0}^{J} b_{jt} x_j^* = X^{*\prime} B_t \tag{C.8}$$

For which B_t is identified under the conditions imposed. Hence, $f_0(0)$ and $f_1(0)$ can be identified through b_{0t} for $t = 0, 1$ and $\tau = f_1(0) - f_0(0)$. □

Proof of Theorem 2.2

Proof. To establish the asymptotic properties of the proposed estimator, it is useful to write the vector of transformed running variable used in the estimation. For unit i, associated with a measurement error group g, let:

$$X_i^* = \mu_g(\tilde{x}_i)\Gamma_i' \tag{C.9}$$

where, $X_i^* = [1, x_{i,1}^*, \ldots, x_{i,J}^*]$, $\mu_g(\tilde{x}_i) = [1, \mu_g^{(1)}(\tilde{x}_i), \ldots, \mu_g^{(J)}(\tilde{x}_i)]$ and $\Gamma_i = L_{J+1} \circ Q$ is the

Hadamard product of the lower diagonal Pascal matrix, L_{J+1}, and the matrix Q_i, where $Q_{(b.c)} = \tilde{x}_i^{c-b}$. For concreteness, if $J = 3$,

$$\Gamma_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1\tilde{x}_i & 1 & 0 & 0 \\ 1\tilde{x}_i^2 & 2\tilde{x}_i & 1 & 0 \\ 1\tilde{x}_i^3 & 3\tilde{x}_i^2 & 3\tilde{x}_i & 1 \end{bmatrix} \tag{C.10}$$

Since $\mu_g(\tilde{x})$ is not observed, but can be consistently estimated from the auxiliary data, let the feasible transformed running variable used in estimation be given by $\hat{X}_i^* = \hat{\mu}_g(\tilde{x}_i)\Gamma_i'$ Then the feasible estimator for the vector $B_+' = [\beta_+, \beta_+^{(1)}, \ldots, \beta_+^{(J)}]$ (and equivalently for B_-), is given by

$$\hat{B}_+ = \left[\hat{X}^{*\prime}\hat{X}^*\right]^{-1} \hat{X}^{*\prime}Y \tag{C.11}$$

and

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[(x_i^* - \hat{x}_i^*)\, B_+ + \varepsilon_i\right] \tag{C.12}$$

$$= \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[(\mu_g(\tilde{x}_i) - \hat{\mu}_g(\tilde{x}_i))\, \Gamma_i' B_+ + \varepsilon_i\right] \tag{C.13}$$

with $\hat{A} = N_p^{-1} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\hat{x}_i^*$ and $\varepsilon = Y - E\left[Y|\tilde{x}, D, G\right]$ Then we can rewrite

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\varepsilon_i + \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime} \left[N_p^{\frac{1}{2}} \left(\mu_g(\tilde{x}_i) - \hat{\mu}_g(\tilde{x}_i)\right)\right] \Gamma_i' B_+\right] \tag{C.14}$$

Then,

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \left(\Gamma_i' B_+ \otimes \hat{x}_i^*\right)' \left[N_p^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i)\right)\right]\right] \tag{C.15}$$

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1} N_p^{-\frac{1}{2}} \sum_{i=1}^{N_p} \hat{x}_i^{*\prime}\varepsilon_i - \hat{A}^{-1} \left[N_p^{-1} \sum_{i=1}^{N_p} \left(\hat{x}_i^{*\prime} \otimes B_+'\Gamma_i\right) \left[N_p^{\frac{1}{2}} \left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i)\right)\right]\right] \tag{C.16}$$

Let h_g be the tuning parameter of a kernel based nonparametric estimator of $\mu_g(\tilde{x}_j)$ such that $h_g \to 0$ and $N_{a,g}h_g \to \infty$ as the sample sizes increase for all g. Finally, let $\lambda_g = \lim_{N_p\to\infty}\left(\frac{N_p}{N_{a,g}h_g}\right)$, for all g.

$$\sqrt{N_p}(\hat{B}_+ - B_+) = \hat{A}^{-1}N_p^{-\frac{1}{2}}\sum_{i=1}^{N_p}\hat{x}_i^{*\prime}\varepsilon_i - \hat{A}^{-1}\left[N_p^{-1}\sum_{i=1}^{N_p}\left(\hat{x}_i^{*\prime}\otimes B_+'\Gamma_i\right)\lambda_g^{\frac{1}{2}}\left[(N_{a,g}h_g)^{\frac{1}{2}}\left(\hat{\mu}_g(\tilde{x}_i) - \mu_g(\tilde{x}_i)\right)\right]\right] \tag{C.17}$$

.

In a mild abuse of notation, let the units that are part of a group g be indexed by a "group unit" denomination l such that $N_p = \sum_{i=1}^{N_p} w_i = \sum_{g=1}^{G}\sum_{l=1}^{N_g} w_{lg}$ for any variable w, with N_g the number of observations in group g in our primary sample.

$$\sqrt{\bar{N_p}}(\hat{B}_+ - B_+) = \hat{A}^{-1}N_p^{-\frac{1}{2}}\sum_{i=1}^{N_p}\hat{x}_i^{*\prime}\varepsilon_i - \hat{A}^{-1}\left[N_p^{-1}\sum_{g=1}^{G}\sum_{l=1}^{N_g}\left(\hat{x}_{lg}^{*\prime}\otimes B_+'\Gamma_{lg}\right)\lambda_g^{\frac{1}{2}}\left[(N_{a,g}h_g)^{\frac{1}{2}}\left(\hat{\mu}_g(\tilde{x}_l) - \mu_g(\tilde{x}_l)\right)\right]\right] \tag{C.18}$$

Let $F_{+,g} = E\left[\left(x_{lg}^{*\prime}\otimes B_+'\Gamma_{lg}\right)\right]$. Also, assume that a CLT holds for the measurement error moment estimator using the auxiliary data such that $\sqrt{N_{a,g}h_g}\left(\hat{\mu}_g(\tilde{x}) - \mu_g(\tilde{x})\right) \to N(0, V_g(\tilde{x}))$. Then the asymptotic variance of $\sqrt{N_p}(\hat{B}_+ - B_+)$ is given by:

$$\Omega_+ = A^{-1}E\left[X^{*\prime}\varepsilon_i\varepsilon_i'X^*\right]A^{-1} + A^{-1}\left[\sum_{g=1}^{G}\lambda_g F_{+,g}'V_g(\tilde{x})F_{+,g}\right]A^{-1} \tag{C.19}$$

Asymptotic normality is achieved by combining the assumption of independence between auxiliary and primary datasets, the asymptotoc normality of $N_p^{-\frac{1}{2}}\sum_{i=1}^{N_p}\hat{x}_i^{*\prime}\varepsilon_i$ and $(N_{a,g}h_g)^{\frac{1}{2}}\left(\hat{\mu}_g(\tilde{x}_j) - \mu_g(\tilde{x}_j)\right)$, and the definitions of λ_g for every $g = 1, ..., G$. Similarly for the left side of the cutoff with Ω_-. Combining the results for both sides of the cutoff, under random sampling gives the result. □

D Two-Sided Fuzzy Designs: Comparison to Davezies and Le Barbanchon (2017)

When considering two-sided fuzzy RD Designs — where the probability of treatment is nonzero and less than one on both sides of the cutoff — with access to auxiliary data, the approach in Davezies and Le Barbanchon (2017) (henceforth D&LeB) is the closest alternative approach to ours in the literature. In this section we illustrate a case in which the direct application

of the D&LeB approach will not recover the intended discontinuity in $E[Y|X]$ at $c = 0$, but our approach (henceforth BBD) can — we thank an anonymous referee for suggesting this comparison of the approaches.

D&LeB consider the two-sided fuzzy RD setting in cases where the researcher has access to information on the true running variable for the treated group. Note, that our approach can be applied in the same settings, however it can also be applied to one-sided fuzzy and sharp designs — as in our leading examples. Importantly, within the two-sided fuzzy setting, there are many measurement error structures for which both approaches are appropriate. Here, we simply highlight an example in which the measurement error structure may be problematic for the direct application of the D&LeB approach. Specifically, when we allow for different measurement error groups — a situation that was not the focus of D&LeB's paper, a discontinuous change in the measurement groups at the cutoff can introduce a spurious discontinuity in the D&LeB estimates. Note that this reflects the case in our leading examples with population distributions changing across counties in the geographic setting and the precision of birth weight measures changing near the cutoff determining treatment for low birth weight babies.

We start by adapting our baseline simulation from section 3 to be a two-sided fuzzy design with a similar first stage relationship between treatment and the true running variable as used in D&LeB. We create a primary data set with 10,000 observations by first drawing the true running variable from a uniform distribution on -1 to 1 ($X \sim U(-1, 1)$). We then specify the following probability of treatment — note that this is the mirror image of the first stage used by D&LeB since their DGP was based on the treatment probability increasing for positive values of X while our baseline simulation has treatment turning on for negative values of X:

$$P(T = 1|X = x) = \frac{3}{8} - \frac{1}{4}\Phi(5x) + \frac{1}{2}\mathbb{1}[x < 0] \tag{D.1}$$

where $\Phi(\cdot)$ is the normal CDF. This first stage relationship, depicted in Figure D.1, implies that the probability of treatment increases from 1/4 to 3/4 when passing the cutoff from above. For the second stage relationship between Y and X, we adapt the DGP from our baseline simulations in two ways. For the two-way fuzzy design we need to derive treated and untreated potential outcomes for all possible values of X. Based on our baseline simulation, we define the following potential outcomes:

$$Y_{i0} = \begin{cases} 0.46 + 1.27x_i + 7.18x_i^2 + 20.21x_i^3 + 21.54x_i^4 + 7.33x_i^5 + v_i & \text{if } x < 1 \\ 0.46 + 0.84x_i - 3.00x_i^2 + 7.99x_i^3 - 9.01x_i^4 + 3.56x_i^5 + v_i & \text{if } x \geq 0 \end{cases} \tag{D.2}$$

$$Y_{i1} = \begin{cases} 0.54 + 1.27x_i + 7.18x_i^2 + 20.21x_i^3 + 21.54x_i^4 + 7.33x_i^5 + v_i & \text{if } x < 1 \\ 0.54 + 0.84x_i - 3.00x_i^2 + 7.99x_i^3 - 9.01x_i^4 + 3.56x_i^5 + v_i & \text{if } x \geq 0 \end{cases} \tag{D.3}$$

where $v_i \sim N(0, 0.1295^2)$ and there is a constant treatment effect equal to $0.54 - 0.46 = 0.08$ which ensures that the expected discontinuity in the reduced form relationship between Y and X is equal to $0.04 = 0.5 \times 0.08$ as in our baseline simulations. The observed outcome is given by as $Y_i = T_i Y_{i1} + (1 - T_i)Y_{i0}$ where the realization of T_i is based on a binomial distribution with probability given by $P(T = 1|X = x)$.

We then generate the measurement error so that there is a discontinuous change in measurement error groups at the cutoff as mentioned above — introducing the problems discussed in Appendix A.2. A simple way to introduce this discontinuity in measurement error groups is to create two measurement groups — one on each side of the cutoff. For this illustration, we create the mismeasured running variable as $\tilde{X}_i = X_i - \epsilon_i$ where the measurement error, ϵ_i is drawn from one of two distributions depending on the X:

1. $U(0, 0.1)$ if $X_i \leq 0$
2. $U(-0.1, 0)$ if $X_i > 0$

Effectively, this generates negative measurement error to the left of the cutoff and positive measurement error to the right. Note that this measurement error structure does not result in the loss of identification issue that is emphasized in D&LeB and highlighted in equation (A.3) in Appendix A since the measurement error does not cross the threshold (i.e. X_i and $\tilde{X}_i$ are on the same side of the cutoff). However, it still creates problems for naive estimators that ignore the group structure of the measurement error.

Finally, we create an auxiliary data set containing the true and mismeasured running variables along with the measurement error group identifier for the treated sample only — thereby matching the setup from D&LeB. Since the expression for $P(T = 1|X = x)$ above implies that the expected probability of treatment is 0.5 for the full sample, the expected sample size for the auxiliary data set is 5000. This auxiliary data for the treatment group is used by both approaches to correct for the measurement error.

Here we will focus on recovering the reduced form discontinuity in $E[Y|X]$ when faced with the mismeasured running variable — where the fuzzy RD estimate would follow naturally after estimating the first stage discontinuity in $E[T|X]$ using the same methods. For BBD, we proceed as outlined in section 3 for our baseline simulation picking the order of the polynomial on each side by the AIC — allowing for up to a seventh order polynomial. For D&LeB, we rely on the authors' Matlab code — available at `https://sites.google.com/site/tlebarbanchon/home`. To implement the D&LeB approach, we must choose the number of nodes for the nonparametric sieve estimator and pick values for several estimation constraints (see D&LeB for a full discussion of the approach). To flexibly model the conditional expectation, we allow for 6 nodes on either side of the cutoff and then follow the example in D&LeB when choosing the estimation constraints.

We conducted a 100 replication Monte Carlo based on the DGP described above. Across the replications, the bias for the D&LeB estimator was -0.0856 while the bias for the BBD estimator was only -0.0029. For illustration, Figure D.2 depicts estimates of $E[Y|X]$ for a single simulation. In addition to the estimates using the BBD and D&LeB approaches, the solid line represents a local linear estimate with Gaussian kernel using the correctly measured running variable to serve as a comparison for the two measurement error corrections. We see that the BBD approach tracks the conditional expectation based on the true running variable quite well while the D&LeB approach deviates substantially — leading to an estimate of the treatment effect that is of the opposite sign and much larger in magnitude than the true treatment effect. Together, these simulation results support the idea that a direct application of the D&LeB approach may struggle when there are discontinuous changes in measurement error structures at the cutoff. While it may be possible to adapt the D&LeB approach to account for this issue by extending their procedure to group specific measurement error structures, this is outside the scope of the current paper.

Figure D.1

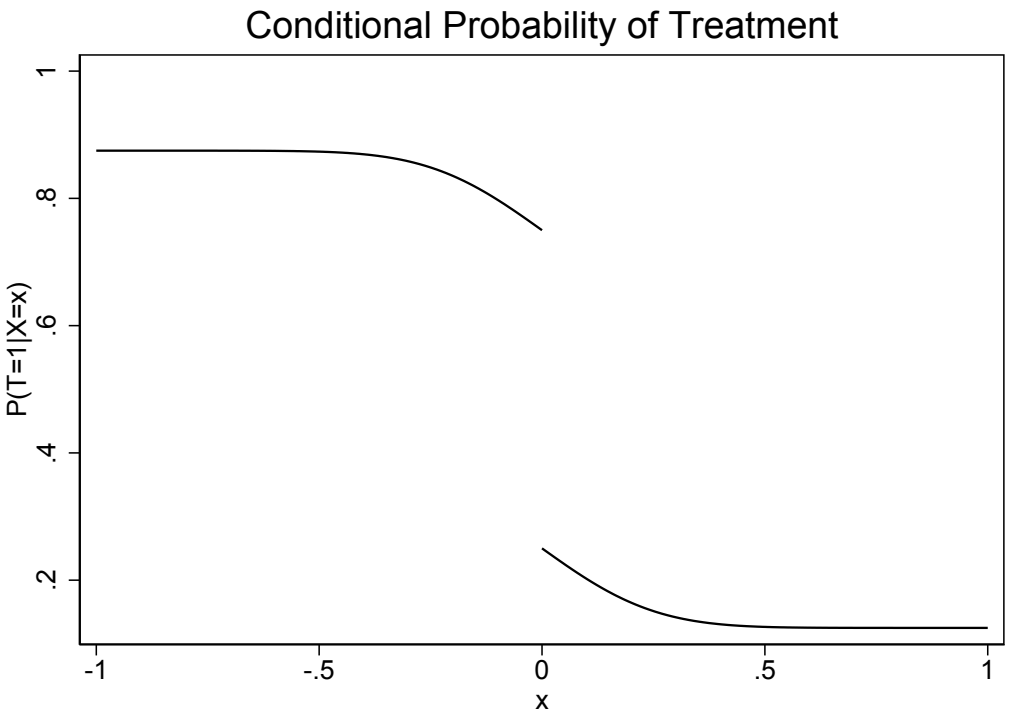


Figure D.2